Physical Fault Tolerance of Nanoelectronics

Thomas Szkopek

Department of Electrical and Computer Engineering, McGill University, Montréal, Québec H3A 2A7, Canada

Vwani P. Roychowdhury

Department of Electrical Engineering, University of California Los Angeles, Los Angeles, California 90095, USA

Dimitri A. Antoniadis

*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, USA*

John N. Damoulakis

Information Sciences Institute, University of Southern California, Marina Del Rey, California 90292, USA
(Received 23 January 2011; published 25 April 2011)

The error rate in complementary transistor circuits is suppressed exponentially in electron number, arising from an intrinsic physical implementation of fault-tolerant error correction. Contrariwise, explicit assembly of gates into the most efficient known fault-tolerant architecture is characterized by a subexponential suppression of error rate with electron number, and incurs significant overhead in wiring and complexity. We conclude that it is more efficient to prevent logical errors with physical fault tolerance than to correct logical errors with fault-tolerant architecture.

Great effort has been devoted to realizing Moore's law [1], whereby the number of components on a single integrated circuit doubles approximately every 2 years. With a greater number of components switching at higher rates, a greater amount of information can be processed per unit of space and time. The most fundamental limit to information processing is set by the spacetime density of physically distinguishable state transitions admitted by quantum mechanics and general relativity [2]. In the more immediate future, computers are unlikely to be very different from contemporary integrated circuits: an essentially planar technology that uses electromagnetic interactions between electrons at room temperature and with energy supplied at an electrochemical potential of ~ 1 eV. Logical gates approaching the molecular scale have been proposed [3], where a bit might be stored on a single electron charge or spin. Limitations are imposed by heat dissipation, wiring, and reliability [4,5]. The logic error rate is presently $\sim 10^{-27}$ per gate operation according to data from the International Technology Roadmap for Semiconductors (ITRS) [6]. Entropy is introduced into the physical state of charge carriers by thermal fluctuations and structural disorder, both of which are increasingly important in devices of reduced physical dimension. The connection between thermodynamics and computation has been well elaborated [7,8], but, nonetheless, the minimum physical system size required for fault-tolerant logic has received comparatively little attention. The related question of minimum system size for storing information has been investigated [9], where the particle number emerges as a natural, dimensionless

size parameter. We likewise adopt the electron number as a dimensionless size parameter.

von Neumann proposed [10] fault-tolerant architectures for computing with faulty components, a concept now extended to triplicate modular redundancy at the system level [11] and gate-level quantum error correction [12–14]. Fault-tolerant architectures are effective in correcting physical errors in nanoscale devices [15–17], but is the fault-tolerant architecture approach optimal for maximizing functional density? It has been postulated that topological excitations of many-body systems provide inherent physical fault tolerance for quantum computation [18,19] in lieu of architectural redundancies. Inspired by this development, we compare the error suppression performance of inherent physical fault tolerance versus architectural fault tolerance for classical computation. The error-rate scaling laws that we derive for complementary transistor logic subject to thermal fluctuations, and ballistic gates subject to atomic placement disorder, are compared in Table I to that of a fault-tolerant architecture.

Complementary transistor logic gates can have an inherently low error rate. CMOS is the most prevalent form of complementary transistor logic, but new materials such as carbon nanotubes [20,21] and semiconductor nanowires [22,23] can also be used for complementary logic. All operate by carrier transport over potential barriers. Physical redundancy and dissipation, associated with entropy removal, stabilizes circuit outputs against the effects of thermal noise and material disorder. The information theoretic origin of robust transistor gate behavior can be explained by considering transistor gate operation in terms

TABLE I. Comparison of physical and architectural error suppression. Logical error-rate approximations for $N \gg 1$ are given.

	Error source	Error per particle	Logical error
Ideal majority vote	$\cdots$	p	$P = \binom{N}{N/2} p^{N/2} \sim (2/\pi N)^{1/2} (4p)^{N/2}$
Subthreshold gate	Thermal $\epsilon = \exp(-\gamma^2 eV/8k_BT)$	$\epsilon^2/4$	$P = \frac{1}{2}\mathrm{erfc}(\sqrt{N\ln[1/\epsilon]}) \sim \frac{1}{2}(2/\pi N \ln[1/\epsilon])^{1/2} \epsilon^N$
Ballistic gate[a]	Disorder $\delta\varphi = -n\pi \cdot \delta r/r$	$\delta\varphi^2/4$	$P \sim (2/\pi N)^{1/2} \delta\varphi^N$
Architecture[b]	Physical gate error	p	$P \le \epsilon_T (p/\epsilon_T)^{N^{\log 2/\log c}}$

[a]Interparticle interaction of the form $V_i \propto 1/r^n$
[b]Triplicate concatenated code, $c = 3$ without error correction ancillae, $c = 9$ with error correction ancillae.

of charge, rather than voltage. A representative, complementary transistor implementation of the universal NAND gate is illustrated in Fig. 1.

Each input charge carrier's presence on an input gate capacitance can be associated with a physical bit value of 1 or 0, while the output is similarly composed of physical bits associated with the presence of charge carriers on the output node. The logical value of an input (output) is represented by the net polarity of the input (output) charge. At the sacrifice of speed, the greatest suppression of error can be achieved with transistors working at subthreshold. The electron-hole conductance will be ideally of the form $G_{n/p} = G_{n_0/p_0} \exp(\pm eV_{GS}/k_BT)$, and the relation between input charge and output charge for the NAND gate is

$$N_{\mathrm{out}} = \frac{N}{2}\frac{G_p - G_n}{G_p + G_n} = -\frac{N}{2}\tanh\left(\frac{N_{\mathrm{in}}}{k_BTC/e^2}\right), \quad (1)$$

where N_{in} is an effective input charge defined by $\exp(-e^2 N_{\mathrm{in}}/k_BTC) = \exp(-e^2 N_A/k_BTC_A) + \exp(-e^2 N_B/k_BTC_B)$, $N/2$ is the charge of a fully polarized node, $G_{n/p}$ is the conductance of the complementary transistor networks, and k_BTC/e^2 is the charge number equivalent of the

FIG. 1 (color online). Universal NAND gate. (a) The output charge N_{out} versus input charges N_A and N_B for a complementary transistor technology. (b) A majority vote physically implemented by competing transistor conductances G_n and G_p suppresses fluctuations in input charge from the output. The entropy of suppressed fluctuations leaves the system through a coupled heat bath, such as substrate phonons.

thermal voltage. The NAND gate suppresses input charge fluctuation from the output charge provided the input fluctuation remains below the noise margin where $|\partial N_{\mathrm{out}}/\partial N_{\mathrm{in}}| < 1$. With the characteristic of Eq. (1), the noise margin is $\gamma \cdot N/2$ where $\gamma = [1 - (k_BT/eV)\times \ln(eV/k_BT)] \sim 0.90$ for a typical operating voltage $V = 1$ V at room temperature. Charge fluctuations below $\sim 0.45N$ are thus suppressed, as compared to the $N/2$ threshold of an ideal majority vote. The probability P that a thermal charge fluctuation $\langle\delta N_{\mathrm{in}}\rangle^2 = k_BTC/e^2$ at the input of a circuit induces a fault at the input to a subsequent circuit will be the probability that the noise margin is exceeded. Considering a fully polarized input, $P = \int_{-\infty}^{-(1-\gamma)(N/2)} p(N_{\mathrm{in}})dN_{\mathrm{in}}$ where $p(N_{\mathrm{in}}) = (2\pi k_BTC/e^2)^{-1/2}\exp(-(1/2)e^2(N_{\mathrm{in}} - N/2)^2/k_BTC)$, evaluated in Table I.

The complementary transistor gate error probability P scales as an ideal majority vote of N samples with error rate $p = \epsilon^2/4 = \exp(-\gamma^2 eV/4k_BT)/4$ per sample. The majority vote is physically implemented by the competition to polarize the output charge per Eq. (1), with electron-hole channel transistors undergoing opposing metal-insulator phase transitions. The same physical fault-tolerance mechanism applies to other complementary transistor gates such as the buffer and more complex logic functions. This suppression of fault probability from input to output is the error correcting process that allows one to construct extremely complex networks of complementary logic without the destructive growth of correlated fluctuations, and thus logic errors, from one circuit to the next. Once operating voltage and temperature are determined, it is the logic element size (i.e., number of charges) that has the most profound effect on suppressing the propagation of fluctuations. The metal-insulator phase transition in the transistor channel is more sharply defined with increasing electron number [24], and the gate is consequently better able to discern the polarization of the input charge.

For device dimensions well below ~ 100 nm, the ballistic regime of device operation is approached. We consider the scaling of error in the representative ballistic system of a single electron spin [25] and more generally $N = 2j$ coupled electrons forming a single magnetic moment with total spin j [26–28]. The spin orientation along a reference axis z can be used to represent a classical logical
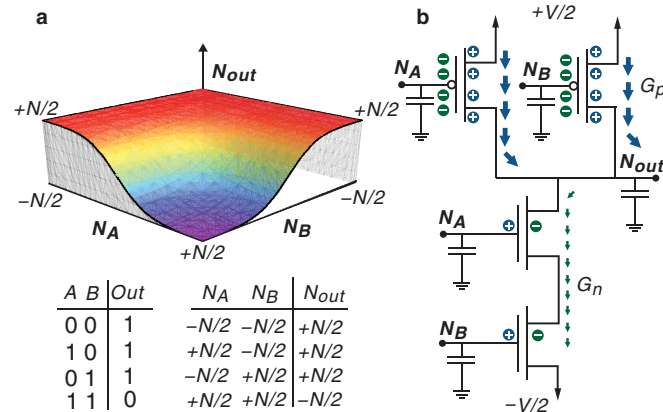
bit. We consider just one way in which errors may arise: the precision of spin placement to a distance δr. Ballistic interaction of two neighboring spins could be implemented by the dipole-dipole interaction $V_i \propto \mu^2/r^3$. The deviation δr in spin position leads to a deviation in the interaction $\delta V_i = -3V\delta r/r$. Unitary evolution by Schrödinger's equation dictates that a spin rotation error $\delta\varphi = -\delta V_i t/\hbar$ will develop over an interaction time t. The time t must be sufficiently long to accumulate at least $\pi = V_i t/\hbar$ radians of spin rotation, corresponding to a classical bit inversion. The spin rotation error incurred in this operation is $\delta\varphi = -3\pi\delta r/r$ radians. With a single electron spin, the probability that the deviation $\delta\varphi \hat{J}_x$ causes an erroneous spin flip is $P = |\langle +1/2| \exp(i\delta\varphi\hat{J}_x)| -1/2\rangle|^2 = \delta\varphi^2/4$. The larger Hilbert space of an $\hbar N/2$ spin than that of a single spin $\hbar/2$ permits a greater tolerance for interaction errors without inducing a reversal of spin polarization. With $N = 2j$ electron spins, the deviation $\delta\varphi\hat{J}_x$ causes an erroneous spin flip with probability $P = \sum_{m>0} |\langle m| \exp(i\delta\varphi\hat{J}_x)| -j\rangle|^2$, approximated in Table I for $N \gg 1$.

Surprisingly, the error rate in the $N = 2j$ electron spin system is equal to that of a majority vote taken on N independent spins with $\delta\varphi^2/4$ error probability per spin. With an error in spin placement of $\delta r = 2.5$ Å and a spin separation as large as $r = 100$ nm, the phase $\delta\varphi = 0.0236 = 1.3°$ and the error probability for a single spin $\delta\varphi^2/4 = 1.4 \times 10^{-4}$. A minimum of $N = 17$ spins would be required to reach an error probability of $P \sim 10^{-27}$. Similar error estimates $P \sim (\delta r/r)^N$ arise for most other interactions, e.g., the Coulomb interaction $V_i = e^2/r$ between electrons in charge based devices. Errors arising from a coupling to a thermal bath can be similarly modeled, where the effective deviation in potential δV_i is ascribed to thermal fluctuations.

We consider now the error scaling of architectural fault tolerance for universal computation. A fault-tolerant architecture is characterized by [13,14] encoding of logical bits into physical bits, detection and correction of physical errors in the encoded logical bits, a universal set of logic operations performed directly on encoded bits. The latter condition implies that only the repetition code is suitable for classical universal computation. The simplest and most efficient architecture is a concatenated repetition encoding with a universal majority gate [29] of Fig. 2. A hierarchy of concatenated encoding is recursively defined, with the final concatenation level L determined by the required error probability. The logical error rate of an architecturally protected bit is given in Table I, where p is the unprotected gate error rate and ϵ_T is a threshold error rate, estimated to be $1/108$ [29,30]. The bound in error rate of Table I can be turned to an equality through numerical refinement of ϵ_T, while preserving the characteristic subexponential scaling of error with physical bit number N [31]. The subexponential scaling of logical error rate with physical bit
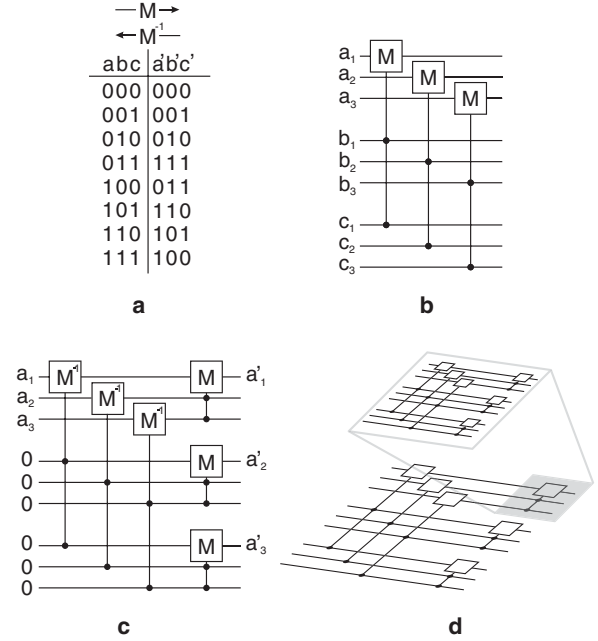

FIG. 2. Fault-tolerant universal majority gate with concatenated triplicate encoding. (a) The majority gate M (and inverse M^{-1}) truth table. (b) Majority gate with inputs and outputs in triplicate. (c) Error correction on a triplicate encoded bit $a = a_1 a_2 a_3$ proceeds by replicating in triplicate with M^{-1}, and taking majority votes to produce the corrected bit $a = a'_1 a'_2 a'_3$. (d) Concatenation recursively applies the triplicate encoding.

number, and thus electron number, distinguishes architectural fault tolerance from physical fault tolerance. The number of physical bits required to protect a level L logical bit is $N = 3^L$ neglecting error correction ancillae, or $N = 9^L$ if error correction ancillae are included.

The logical error rates for physical and architectural fault-tolerant schemes are compared in Fig. 3 for device operating voltages anticipated by the ITRS [6]. The architectural model was applied to physical gates with thermally limited error rate with $eV = 0.97$ eV, $k_B T = 26$ meV and thus $p = (1/2)\text{erfc}[\exp(-\gamma^2 eV/8k_B T)] = 2.9 \times 10^{-3}$, and ballistic single electron gate limited by atomic disorder $\delta r/r = 2.5$ Å/100 nm in Coulomb interaction ($V_i \propto 1/r$) with corresponding error rate $p = \delta\varphi^2/4 = 1.5 \times 10^{-5}$. The physical fault tolerance intrinsic to subthreshold complementary transistors gives superior error suppression per electron number as compared to architectural fault tolerance. In short, constructing a fault-tolerant majority gate from faulty majority gates is not as efficient as majority voting implemented directly at the physical level in subthreshold complementary logic. Moreover, the overhead associated with architectural fault tolerance can be significant, particularly the wiring for addressing individual bits. We also conclude that $N = 16$–24 electrons are minimally required to suppress thermally induced errors below the error rate of $P \sim 10^{-27}$ at room temperature. Sources of
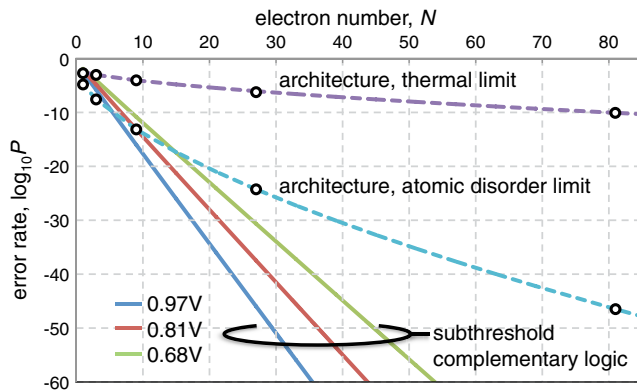
FIG. 3 (color online). Logical error-rate versus electron number. The thermal fluctuation limit ($k_B T = 26$ meV) of complementary logic error rates is compared against that of architectural fault tolerance with physical majority gates limited by thermal fluctuation or atomic disorder in the ballistic limit. A concatenated triplicate code is assumed, with nonancilla electron number $N = 3^L$ highlighted by **o**.

additional charge fluctuation such as elevated operating temperature, charged defects, and parasitic signal coupling will require an increased electron number. Similar scaling of logical error with particle number will apply to any logic element with a single particle error rate of order $\sim \exp(-eV/4k_B T)/4$ resulting from thermal excitation across a potential barrier.

Universal computation thus appears to abide by the adage "a stitch in time saves nine," where it is more efficient to prevent errors than to correct errors. Since subthreshold logic effectively implements an ideal majority vote, the question naturally arises as to whether a more efficient architectural fault-tolerant scheme approaching the ideal majority vote can be devised. The answer notwithstanding, fabrication imperfections render the thermal limit difficult to achieve, and the question remains as to the physical limits to structural order, such as dopant atom location, for a system with sufficient structural complexity to permit universal computation.

[1] G. Moore, Electronics **38**, 114 (1965).
[2] S. Lloyd, Nature (London) **406**, 1047 (2000).
[3] K. Moth-Poulsen and T. Bjornholm, Nature Nanotech. **4**, 551 (2009).
[4] V. V. Zhirnov, R. K. Cavin III, J. A. Hutchby, and G. I. Bourianoff, Proc. IEEE **91**, 1934 (2003).
[5] J. D. Meindl, Q. Chen, and J. Davis, Science **293**, 2044 (2001).
[6] International Technology Roadmap for Semiconductors, http://public.itrs.net (2005–2009).
[7] C. H. Bennett, Int. J. Theor. Phys. **21**, 905 (1982).
[8] R. Landauer, IBM J. Res. Dev. **5**, 183 (1961).
[9] R. M. Swanson and J. D. Meindl, IEEE J. Solid-State Circuits **7**, 146 (1972).
[10] J. von Neumann, *Automata Studies*, Annals of Mathematics Studies No. 34 (Princeton University Press, Princeton, NJ, 1956), p. 43.
[11] L. Sterpone, *Electronic System Design Techniques for Safety Critical Applications* (Springer, New York, 2008).
[12] P. Shor, in *Proceedings of the 37th Annual Symposium on Foundations of Computer Science* (IEEE Computer Society Press, Washington, DC, 1996).
[13] D. Gottesman, Phys. Rev. A **57**, 127 (1998).
[14] J. Preskill, Proc. R. Soc. A **454**, 385 (1998).
[15] J. R. Heath, P. J. Kuekes, G. S. Snider, and R. S. Williams, Science **280**, 1716 (1998).
[16] R. M. Roth, W. Robinett, P. J. Kuekes, G. S. Snider, and R. S. Williams, Nanotechnology **20**, 135201 (2009).
[17] M. Haselman and S. Hauck, Proc. IEEE **98**, 11 (2010).
[18] M. H. Freedman, A. Kitaev, M. J. Larsen, and Z. H. Wang, Bull. Am. Math. Soc. **40**, 31 (2003).
[19] C. Nayak, S. H. Simon, A. Stern, M. Freedman, and S. D. Sarma, Rev. Mod. Phys. **80**, 1083 (2008).
[20] A. Javey, Q. Wang, A. Ural, Y. M. Li, and H. J. Dai, Nano Lett. **2**, 929 (2002).
[21] P. Avouris, R. Martel, V. Derycke, and J. Appenzeller, Physica B (Amsterdam) **323**, 6 (2002).
[22] Y. Huang, X. F. Duan, Y. Cui, L. J. Lauhon, K. H. Kim, and C. M. Lieber, Science **294**, 1313 (2001).
[23] D. Wang, B. A. Sheriff, and J. R. Heath, Small **2**, 1153 (2006).
[24] C. Borgs, J. T. Chayes, H. Kesten, and J. Spencer, Commun. Math. Phys. **224**, 153 (2001).
[25] D. Loss and D. P. DiVincenzo, Phys. Rev. A **57**, 120 (1998).
[26] A. Barra, A. Caneschi, A. Cornia, F. de Biani, D. Gatteschi, C. Sangregorio, R. Sessoli, L. Sorace, V. Barone, O. Hod, and G. Scuseria, J. Am. Chem. Soc. **121**, 5302 (1999).
[27] C. Milios, A. Vinslava, W. Wernsdorfer, S. Moggach, S. Parsons, S. Perlepes, G. Christou, and E. Brechin, J. Am. Chem. Soc. **129**, 2754 (2007).
[28] M. Mannini, F. Pineider, P. Sainctavit, C. Danieli, E. Otero, C. Sciancalepore, A. Talarico, M.-A. Arrio, A. Cornia, D. Gatteschi, and R. Sessoli, Nature Mater. **8**, 194 (2009).
[29] P. O. Boykin and V. P. Roychowdhury, in *Proceedings of the International Conference on Dependable Systems and Networks 2005* (IEEE Computer Society Press, Washington, DC, 2005).
[30] K. Nikolić, A. Sadek, and M. Forshaw, Nanotechnology **13**, 357 (2002).
[31] P. Aliferis, D. Gottesman, and J. Preskill, Quantum Inf. Comput. **6**, 97 (2006).