R. W. Keyes
R. Landauer

Minimal Energy Dissipation in Logic

Abstract: Minimal energy dissipations for the logic process based on thermodynamics and general phase space considerations are known. The actual availability of these minimal dissipations has not, however, been demonstrated. These minimal dissipation sources in a computing system also act as noise sources and thereby lead to questions about the ultimate available reliability of the computing process. A new and hypothetical device is presented in this paper and used to construct a physically analyzable computing system. It is demonstrated that this system has dissipations larger than, but of the same order of magnitude as, the original minimal quantities. It is also shown that any required reliability can be obtained with this device, without increased energy expenditure, but at the expense of an increasing time per computational step.

Introduction

The past decade has brought a growing realization that the processing of information, whether carried out in computers, in biological systems, or with paper and pencil, requires the use of real physical degrees of freedom, subject to the laws of physics. Studies of the ultimate physical limitations of information handling, even though they are still in a very rudimentary state, constitute the beginning of a genuine physical science of epistemology. This point has been made particularly eloquently by Lederberg in a recent syndicated newspaper column.[1] The work in this general field of fundamental computer limitations has been reviewed by Freiser and Marcus.[2] A book chapter by one of the authors[3] is awaiting publication.

A central question in this area has been, Is there a minimal energy dissipation associated with the nonlinear processes that carry out the typical logic in a computer? The association of an amount $\frac{1}{2}kT$ of random thermal energy with a degree of freedom has always made it plausible that the intentional logic signals must be associated with a comparable energy. This was understood by von Neumann, apparently as early as 1949.[4] Von Neumann indeed suggested that an energy kT is dissipated "... per elementary act of information, that is, per elementary decision of a two way alternative and per elementary transmittal of 1 unit of information." A more exact understanding of the reason for the amount of the dissipation was provided by one of the authors,[5] who pointed out that general purpose digital computers require the capability to throw away information, and that it is these information reduction processes which, in turn, require an energy dissipation of order kT per logical step. In particular, for example, in an operation that requires a bit to be reset to ONE, regardless of its initial value, and in which the data flow initially gives ZERO and ONE equal probabilities, the minimal energy dissipation is $kT \log_e 2$.

Dissipation is inevitably associated with thermally induced fluctuations, as is made clear by the existence of fluctuation-dissipation theorems. Therefore, if computers must have dissipation, they must also have internal noise sources. A very basic and important question then becomes, Is computing to an arbitrary reliability specification possible, or is there an irreducible error probability? This paper demonstrates that, *at least as far as general statistical mechanical considerations are concerned*, there are no obvious reliability limitations to the computing process. A given computer has a nonvanishing error probability at each step, but a more reliable computer can be built upon demand. Our answer, however, is in no sense a final settlement of the reliability problem. To make more reliable computers, we invoke the physical availability of potentials, whose actual realizability is not terribly clear, and which, even if realizable, come with uncertain implications about the size of the computer, both geometrically and in terms of the number of particles required in it. All these subsidiary considerations may (and are perhaps even likely to) negate our purely statistical mechanical conclusion that arbitrarily accurate information processing can be realized. Furthermore our conclusions apply only to very slow computing processes. While we show that

The authors are located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

increasing reliability does not require increasing energy, we do have to allow an unbounded time scale for the process *at each step*.

Energy of switching

This paper is, in addition, motivated by the fact that a more recent paper by Neyman[6] concludes that reliable switching processes require larger dissipations than are indicated in Landauer's original discussion,[5] with the energy dissipation increasing to larger and larger amounts as the reliability specifications become more stringent. Since Landauer's discussion invoked the second law of thermodynamics to find a lower bound for energy dissipations, and did not demonstrate the actual achievability of the minimal energies, Neyman's results must be considered as plausible and deserve thoughtful consideration. We shall point out that Neyman's higher energies, required for reliable switching, are in fact not needed; smaller amounts comparable to those in the original thermodynamic analysis are achievable, if we are satisfied with *very slow* switching.

Unfortunately, while the slow switching is proceeding, unswitched information held elsewhere in the computer may be deteriorating unintentionally unless special measures are taken to prevent this adverse effect. This point was discussed in Landauer's original paper,[5] in which a specific model was analyzed [in Eq. (5.4) of that paper] to show that long computations in systems with many elements require energy dissipation appreciably greater than kT. In this paper, however, we demonstrate that the model leading to Eq. (5.4) of Landauer's original paper[5] is unnecessarily pessimistic, just as Neyman's equations are.

Neyman starts from an entropy increase per switching event,

$$\Delta S > k \log_e (r\Delta I). \tag{1}$$

Here k is Boltzmann's constant, I is the information and $r = 1/p_1$, where p_1 is the probability of an error due to thermal fluctuations in an "individual measurement." There is some uncertainty in both the definition of ΔI and the origin of Neyman's equation. Equation (1) has a close formal relationship to Eq. (14.31) of Brillouin's book[7] (originally pointed out to the authors by D. W. Jepsen) and to a very similar discussion by Ligomenides.[8] These authors, however, are concerned with measurements of energy in an harmonic oscillator potential, which is not an obvious model of a bistable computer element. Equation (1) leads Neyman to an energy loss

$$\Delta E \geq kT \log_e (r\Delta I). \tag{2}$$

Consider the bistable well shown in Fig. 1 and used as a model in Ref. 5 and in a subsequently published elaboration.[9] Switching proceeds by tipping the well;

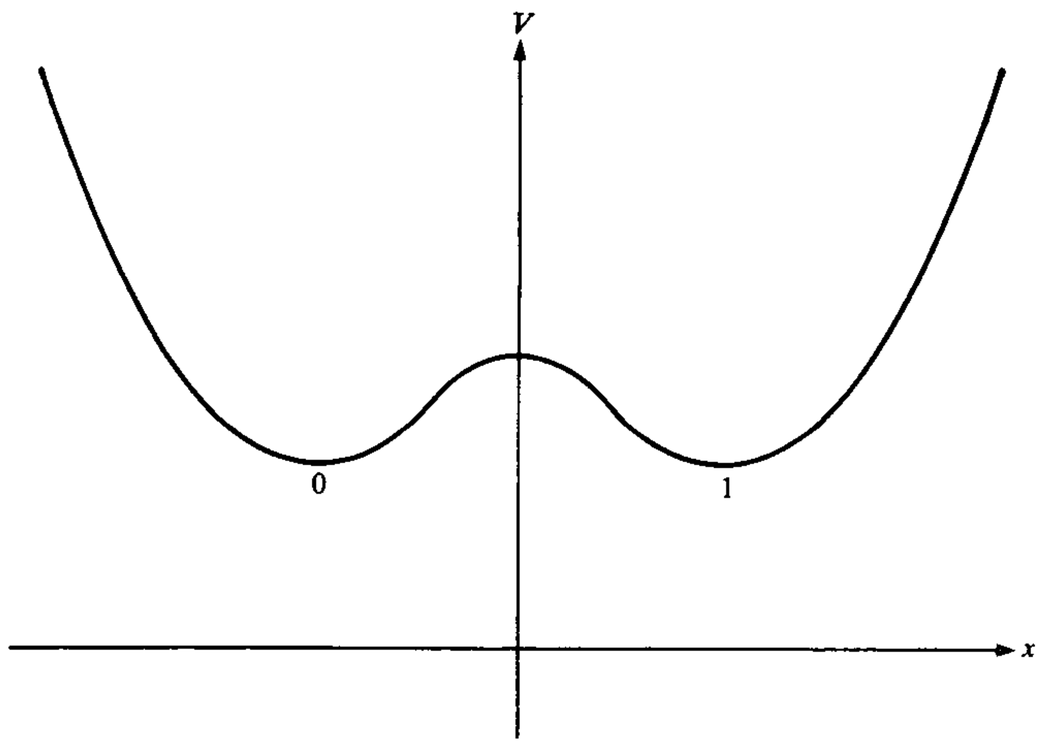

Figure 1 Bistable potential well; x is a generalized coordinate representing the quantity that is switched.

thus if the right-hand well is favored, the particle will end up there, regardless of its initial position. Now if we want to be very sure that switching has taken place, we must bias the well by many times kT. Indeed, the smaller p_1 is to be, the larger the required energy difference between the wells must be. If this energy difference is dissipated, then Eq. (2) seems very plausible. We shall, however, show that for the bistable well of Fig. 1 a dissipation $kT \log_e 2$ is in fact achievable in the RESTORE-TO-ONE operation, even for very small values of p_1.

Small p_1, or accurate switching, is indeed associated with large biasing forces, but these forces are not necessarily associated with large energy losses. The basic physical point involved can be made as follows. A modest biasing force is adequate to give a reasonably high probability that the particle is located in the favored well; if an additional biasing force is applied subsequently, there is a high probability that the particle will not be subject to further well jumping. Hence the additional force has a small probability of causing additional energy dissipation. Thus if the biasing force is increased slowly compared with the switching time, most of the energy dissipation should occur in the early portion of the bias application. By contrast, if the bias is applied quickly greater energy dissipation results, accompanied by faster switching as discussed in connection with Eq. (5.4) of Ref. 5.

To make this discussion more quantitative consider an ensemble of wells with the ZERO and ONE states initially populated with equal probability. Let these wells be subjected to a slowly increasing energy bias U. The fraction η of wells in the favored state is

$$\eta = [1 + \exp(-U/kT)]^{-1}. \tag{3}$$

In going from bias energy U to $U + dU$, a fraction $d\eta$ of wells shift into the favored state. Each of these gives up

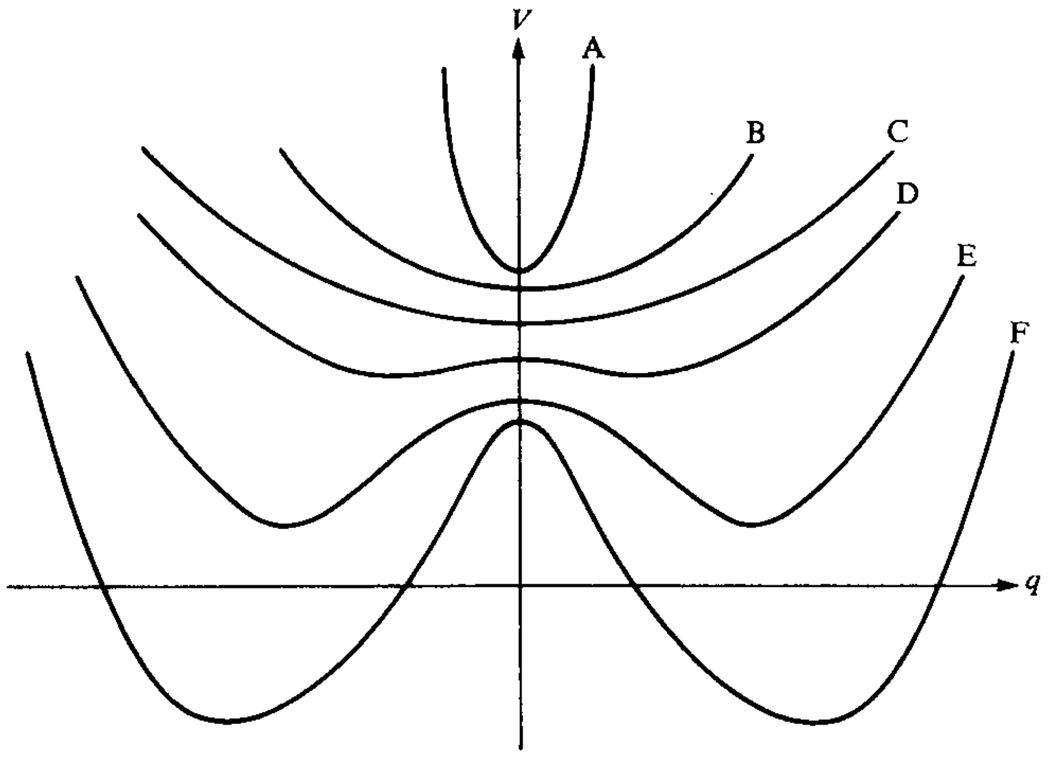

Figure 2 Time sequence of potentials starting at A (for a particle known to be near $q = 0$) and changing continuously to the deep bistable wells at F.

energy U in the process. The total dissipation per well as U is increased is therefore

$$\int_{\frac{1}{2}}^{\eta} U d\eta = kT[\log_e 2 + \eta \log_e \eta + (1 - \eta) \log_e (1 - \eta)], \quad (4)$$

as can be shown by an elementary integration. Note that the form $Ud\eta$ is closely related to HdM or EdP in electrically activated storage systems such as magnetic cores. The energy dissipation in Eq. (4) follows exactly the decrease in entropy. In particular, when $\eta = 1$ an energy $kT \log_e 2$ is dissipated, as was originally derived from entropy considerations. It should be understood, however, that in Eq. (4) we have calculated the *energy* changes of the particle, not the *free energy* changes. It is true that this energy is given up by the particle and delivered as heat to the surroundings, just as when a set of spins is oriented into a favored direction. As in the case of spins, however, if we let the system subsequently randomize again, the heat is resorbed by the information-bearing degree of freedom under consideration. To let Eq. (4) represent a permanent dissipation, we cannot afford to let the system re-randomize. After removing the bias the system must be subject to its next use, before it acts as an "adiabatic demagnetization" refrigerator. This, however, is in accord with common sense. One does not, in a useful computing system, expect to have jumping from ZERO to ONE on the basis of random thermal agitation. However, for one to follow a system through a sequence of operations and study the interactions among wells, this one-well model is inadequate and a more detailed model of a computer has to be examined.

As was suggested previously, the minimal dissipation is available only for a slow switching process, one that is slow compared with the relaxation time of the undisturbed system. This condition means that we lose information in undisturbed locations while switching is going on elsewhere. In subsequent sections, however, we argue that the burden of retaining information can be put on wells that are deeper and therefore more protected against information loss than the wells that are being tipped.

We now begin a discussion of a more complex and realistic computer system capable of performing all logical functions and not confined to the relatively trivial case of a one-time RESTORE-to-ONE operation. The basic point of our discussion as mentioned previously is that information not subject to switching can be protected by higher barriers against deterioration.

Description of the computing system

The computation scheme we have in mind was essentially invented by the late von Neumann and is described, for example, by Wigington.[10] It is a method designed to use systems which, under external control, can be taken continuously from a monostable state into bistability and back to monostability in a cyclic fashion. Von Neumann applied his invention to parametric excitation in tuned circuits with nonlinear reactances. A similar approach has been invoked for the utilization of tunnel diodes by Goto.[11] We invoke here a logic scheme based on the same notions. Our physical device, however, is a particle in a potential well. The potential well is modulated periodically in time, with the modulation converting it from a well with a single-minimum to a well with two minima, and back again. A sequence of successive well shapes going from the single-depression well to the double well is illustrated in Fig. 2. To the solid state physicist these curves are reminiscent of the temperature dependence of the free-energy curves for a ferroelectric going through a second-order transition. In analogy with that phenomenon we call the well states near curve C, i.e., near the borderline between monostability and bistability, the "soft" states. Figure 3 illustrates a related time-dependent sequence of wells, the significance of which is developed in the subsequent text.

The scheme envisions that each of the logic stages in a computer is associated with a time-modulated well as shown in Fig. 2 and that the stages are grouped into different "phases"; within a phase group the stages are at the same part of their excursions through the well shapes of Fig. 2. In analogy with von Neumann's scheme we call the source of time variation in the potential the "pump." In our case, the pump could be a series of suitably chosen charges that are brought cyclically toward and away from a charged particle whose coordinate q is used to represent the information. The particles in the flat-bottomed soft state are particularly susceptible to external influences.

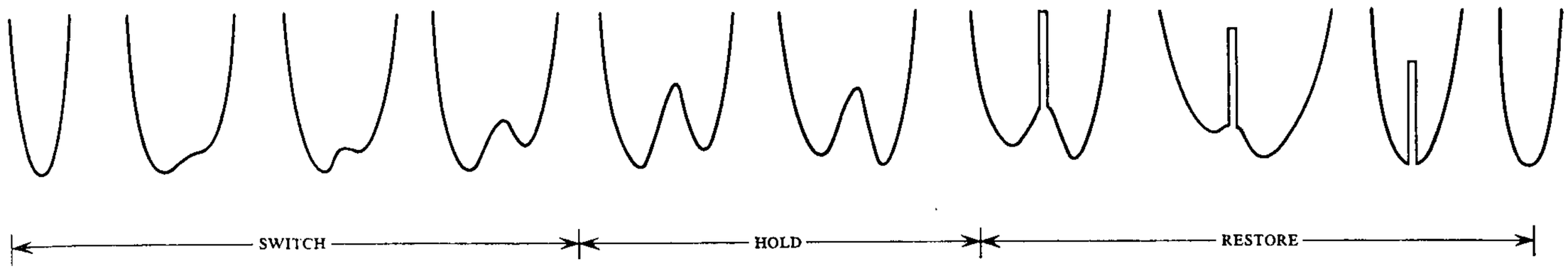

Figure 3 The sequence of shapes through which a well coupled to preceding and following stages passes during a complete period of the cycle shown in Table 1. During SWITCH the well passes successively through the stages A to F shown in Fig. 2; the asymmetry is the result of the coupling to the previous stage. During HOLD the information in the well influences the switching of the subsequent stage. Small variations in shape take place during the HOLD stage due to the changes in the forces exerted on the well by the previous stage, which is being restored, and the following stage, which is being switched. During RESTORE the well is biased by coupling to wells of the subsequent phase. The thin barrier at the center of the well during the RESTORE cycle prevents excessive dissipation due to coupling backwards from the following stage.

The overall scheme uses this property; each particle is loosely coupled to particles that belong to a phase more advanced in time, so that as a particle reaches the soft states it is pushed one way or another by the influence of other particles that are already stabilized in the deep bistable potential near state F of Fig. 2.

A more detailed description of the sequence of events through which our wells are taken is provided in Table 1. Each of the bistable elements belongs to one of at least three phases. That is, the processor is driven cyclically and there are elements at three different parts of the cycle. The stages at different parts of the cycle are designated α, β and γ in Table 1; there are many elements in each phase, which may be designated $\alpha_1, \alpha_2, \cdots, \beta_1, \beta_2$, etc., when necessary.

We start with a β-phase well about to begin the SWITCH part of its cycle (at the left-hand end of the second row in Table 1). Such a well is in the deep, single-trough configuration shown at the top of Fig. 2 and labeled A. The value of the switchable coordinate q of the particle is very close to zero. The particle is tied through springs to three (or another odd number of) other particles belonging to the preceding phase α that is at F in its cycle. Each of the springs exerts a force of the form $\pm Aq_F$ on the particle, where q_F is the magnitude of the displacement of the minima at stage F. Thus the total force exerted by the springs acts in a direction determined by the majority of the wells of the preceding stage. This is, in von Neumann's words, "majority logic." As the cycle progresses the well passes through its soft stages B, C and D of Fig. 2, in which it is easily influenced by the springs and is displaced in the majority direction. Then, as the well changes from its soft state towards E, the particle becomes locked firmly in the direction to which it was initially influenced while in the soft state.

References 10 and 11 show that this majority logic, if combined with negations, can carry out all logical functions. Negation is carried out by simply coupling a particle in the earlier phase to a particle in the succeeding

Table 1 Sequence of events in the three phase groups α, β and γ. In HOLD an element is near the deep-well state F; during SWITCH the element changes from A to F under the influence of elements in the HOLD stage; during RESTORE an element is returned from F to A to prepare it for a new cycle.

	Time			
Stage	*1*	*2*	*3*	*1*
α	HOLD ↓	RESTORE	SWITCH →	HOLD
β	SWITCH →	HOLD ↓	RESTORE	SWITCH
γ	RESTORE	SWITCH →	HOLD ↓	RESTORE
α	HOLD	RESTORE	SWITCH →	HOLD

phase through a potential that reaches a minimum when one well is in the ZERO state and the other well is in the ONE state. Such an interaction can be visualized by thinking of the spring as acting through a lever, pivoted at its center, that reverses the direction of the force.

An AND operation, for example, which gives a ONE output only if its two inputs are both at ONE, is performed by having as the three influencing inputs the two variables on which the AND is to be performed and a third particle kept permanently in the ZERO state.

Let us continue consideration of the β well's being influenced and SWITCHed. The well we are following can be deformed sufficiently slowly that the particle is at all times arbitrarily close to equilibrium between the left- and the right-hand pockets. Such a slow deformation must be continued until the energy difference between the two pockets attains a value, say V_1, associated with curve E in Fig. 2, that gives some large desired Boltzmann distribution ratio between the pockets and ensures that the particle is where we want it. Here "slow" is a relative term. In fact, the deformation must be so slow that particles can equilibrate across the barrier V_1' that exists when the

potential *difference* reaches V_1, but it cannot be so slow that the particles holding the information in the preceding wells leak over or through their barriers. This balance can be arranged by making the barrier in the completely switched α well have some very high value V_2', associated with the deepest potential F in Fig. 2, through which passage of a particle is so slow that the time to attain the potential difference V_1 in the well being switched may be considered to be very small in comparison.

At E we have arrived at a state in which the particle has a satisfactorily large probability exp (qV_1/kT) of being in the desired well. No irreversibility has occurred up to now; if we reverse the sequence of potentials, we will retrace our sequence of particle distributions. The avoidance of irreversibility is discussed in more detail in later paragraphs. Now we wish to stabilize and standardize this condition so that the information contained in this stage can be used to influence the following stage. Stabilization can be effected by rapidly raising the barrier between the two halves of the potential to the necessary large value V_2'.

The barrier can be erected on top of the central potential maximum or, alternatively as shown in Fig. 2, by further deepening of the wells from E to F. If the latter is done, we must carry out the further deepening quickly compared with the inter-well relaxation times associated with the deep wells near F; otherwise our operation takes so long that we lose information in the process. At the same time, we must go slowly compared with the relaxation time for redistribution within a well so that the change of well shape, within the favored well, does not generate appreciable dissipation. Part of the relaxation process within the favored well relates to the disappearance of the bias exerted from the preceding phase. This bias has a magnitude which depends on how much agreement was involved in the majority poll, but we need a distribution in the state F that is independent of history in order to exert a standardized force on the following stages. In any case, however, the use of sufficiently sharp well minima at F will ensure the history-independent distribution.

The events described so far constitute one-third of a cycle. During the next third, the HOLD part of the cycle, the β well influences the succeeding stage γ and is assumed to be so firmly stabilized by the barrier V_2 that no dissipation is caused by the reaction of this succeeding stage.

The last third of the cycle RESTORES the potential to its original state A (Fig. 2). Let us first consider what would happen if we simply changed the potential from F back to A through the sequence of Fig. 2. As the bistable potential of the β wells is diminished, the β wells become susceptible to the influence of the γ well. A γ well matches the majority disposition among the β wells to which it is connected, but does not necessarily match each of the β wells, and therefore tends to induce some transitions in the β wells. If these occur at some relatively deep potential, such as at E where the minima are far apart, and the spring coupling can exert an appreciable bias, then large dissipative losses can occur. We prevent this by erecting a thin barrier on top of the potential at F, before commencing to diminish the barrier. The sequence of well shapes, including this barrier, is illustrated in Fig. 3.

The thin barrier prevents inter-well transitions and maintains a sequence of quasi-equilibrium distributions. Finally, when the β potential has returned to the A stage, the two halves are not far apart and the bias supplied by the springs is modest. (The narrower the potential well, the smaller the bias.) After the barrier is removed, only a diffusive force remains to erase the memory of the original β states. Indeed, if we want to be sure that the bias is at a minimum, we could retain the thin β barriers until the γ well has also returned to its narrow monostable potential configuration. In the final diffusion process, after removal of the thin barrier, we change from a distribution that fills one half of a well to one that fills the whole well. Here an unavoidable nonequilibrium process finally occurs. The entropy increases by $k \log_e 2$ and the free energy decreases by $kT \log_e 2$, with the loss in free energy turning up as irreversible heating.

Thus we have found an irreversible heat generation of $kT \log_e 2$ per input variable and per logical step. This amount generally exceeds the minimal amount predicted by the earlier theory,[5] but is of the same order of magnitude. Thus the AND operation has two inputs and a loss of $2kT \log_e 2$, or $1.386\ kT$ per operation. The earlier theory predicted the smaller energy loss $0.75kT \log_e 3$, or $0.824\ kT$. The most pronounced difference is in the simple process of information transfer, unaccompanied by any logic operation. No logical irreversibility is involved and, therefore, no minimal heat generation according to the earlier theory; the specific method proposed here generates $kT \log_e 2$ of heat. These differences may simply mean that the presently proposed method is not the optimal invention, and that there are less dissipative methods available. Alternatively they could mean that the minima obtained from general phase space considerations are not really available in a general purpose computer. (Note, however, that elimination of the thin barriers on the RESTORE cycle, which results in much larger dissipations for some logic functions, would leave the information transfer function dissipationless.)

It has not been necessary to make any compromise with reliability to achieve this energy dissipation. Any desired reliability can be attained by properly choosing the potential difference V_1 that controls the Boltzmann factor, by deforming the wells so slowly that equilibrium can be attained across the barrier V_1' that exists at this stage, and by making the stabilizing barrier V_2' large enough to prevent any deterioration of the information in the well

while it is contributing to the control of the next stage. The barrier V_2' must be increased if V_1' is increased, because the time needed to attain equilibrium across V_1' increases.

Discussion

Note that it is necessary to have three different pump phases in the system. Two pump phases are inadequate; in that case a well that has passed on its information, and has been returned to its single-well state, is coupled simultaneously to the wells that have received information and to wells that are about to transmit information, and is influenced equally by them. If we allow three pump phases, then a well about to be influenced can be coupled simultaneously to wells that are the source of the influence and to wells that have been influenced but have already been restored to their neutral single-well state. Three different pump phases are accepted as adequate.[10,11]

We now elaborate on a result used above in discussing the energetics: As we change with time through a series of potential distributions and let the particle come into thermal equilibrium with each potential, the irreversible dissipation can be minimized to any prescribed extent by going through the sequence sufficiently slowly. The adjustment of a particle distribution to a narrowing or widening potential well is like, for example, the isothermal compression or expansion of a gas; it may involve heat flow but need not generate any irreversible heating if done slowly. If in changing the potential we move our particle a distance x in a time t in a medium of mobility μ, we have to change potentials slowly enough to keep loss terms of the type $x^2/\mu t$ small compared with the essential losses to be identified otherwise. Or, alternatively, if we think of our particle in a medium of mobility μ, the rate of entropy generation in an interval dx is

$$\partial S/\partial t = \mathbf{j}\cdot\nabla F\, dx = (\mathbf{j}^2/\mu)\, dx, \tag{5}$$

where F is the free energy (or chemical potential since we are dealing with one-particle systems) and $\mathbf{j}$ is the particle current flow. In thermal equilibrium, i.e., if the Boltzmann distribution is obeyed, ∇F vanishes. If the potential shape is now changed slightly, we can always select the origin of the new potential such that the new free energy, after equilibration, equals the original free energy. The redistribution of particles is accompanied by an irreversible entropy generation, expressed by Eq. (5), but this quantity can be minimized to any required extent by changing the potential so slowly that the relative population changes that occur within a relaxation time are kept sufficiently small. The irreversibility will arise, therefore, in those parts of the cycle in which we are mapping particle distributions into each other, which are not in equilibrium; at those points the free energy *has* to change.

Conclusions

As far as statistical mechanics goes, the effect of thermal fluctuations can be made as small as desired by invoking suitably chosen potentials without creating large energy dissipation, but this is accomplished at the expense of the speed of the computation. These considerations, however, do not assure us that the required potentials are physically realizable. Furthermore, slow switching can be considered to create a small bandwidth for the noise power; it is then not too surprising that the noise effects are limited. The energy-reliability tradeoff for faster speeds remains unclear at this point.

References

1. J. Lederberg, *San Francisco Examiner,* January 5, 1969.
2. M. J. Freiser and P. M. Marcus, *IEEE Trans. Magnetics* **Mag-5,** 82 (1969).
3. R. Landauer, *Proceedings of the Conference on Fluctuation Phenomena in Classical and Quantum Systems,* Chania, Crete, Greece, August 1969, edited by E. D. Haidemenakis, Gordon and Breach, Science Publishers, Inc., New York 1970.
4. J. von Neumann, *Theory of Self-Reproducing Automata,* University of Illinois Press, Urbana 1966, p. 67; *The Computer and the Brain,* Yale University Press, New Haven, Conn. 1958.
5. R. Landauer, *IBM J. Res. Develop.* **5,** 183 (1961).
6. M. S. Neyman, *Telecommunications and Radio Engineering* **21,** 68 (1966).
7. L. Brillouin, *Science and Information Theory,* 2nd edition, Academic Press, Inc., New York 1962.
8. P. A. Ligomenides, *IEEE Spectrum* **4,** 65 (1967); *Proc. I.R.E.E. Australia* **29,** 65 (1968).
9. R. Landauer, *IEEE Spectrum* **4,** 105 (1967).
10. R. L. Wigington, *Proc. IRE* **47,** 516 (1959).
11. E. Goto, K. Murata, K. Nakazawa, K. Nakagawa, T. Motsoka, Y. Motsuoka, Y. Ishibashi, H. Ishida, T. Soma and E. Wade, *IRE Trans. Electronic Computers* **EC-9,** 25 (1960).

Received October 13, 1969