
PERGAMON

Studies in History and Philosophy of Modern Physics 34 (2003) 501–510

Studies in History and Philosophy of Modern Physics

www.elsevier.com/locate/shpsb

Notes on Landauer's principle, reversible computation, and Maxwell's Demon

Charles H. Bennett

IBM Research Division, Yorktown Heights, NY 10598, USA

Abstract

Landauer's principle, often regarded as the basic principle of the thermodynamics of information processing, holds that any logically irreversible manipulation of information, such as the erasure of a bit or the merging of two computation paths, must be accompanied by a corresponding entropy increase in non-information-bearing degrees of freedom of the information-processing apparatus or its environment. Conversely, it is generally accepted that any logically reversible transformation of information can in principle be accomplished by an appropriate physical mechanism operating in a thermodynamically reversible fashion. These notions have sometimes been criticized either as being false, or as being trivial and obvious, and therefore unhelpful for purposes such as explaining why Maxwell's Demon cannot violate the second law of thermodynamics. Here I attempt to refute some of the arguments against Landauer's principle, while arguing that although in a sense it is indeed a straightforward consequence or restatement of the Second Law, it still has considerable pedagogic and explanatory power, especially in the context of other influential ideas in nineteenth and twentieth century physics. Similar arguments have been given by Jeffrey Bub (2002).

Keywords: Landauer's principle; Thermodynamics of information processing; Maxwell's demon; Second law of thermodynamics

1. Landauer's principle

In his classic paper, Rolf Landauer (1961) attempted to apply thermodynamic reasoning to digital computers. Paralleling the fruitful distinction in statistical physics between macroscopic and microscopic degrees of freedom, he noted that some of a computer's degrees of freedom are used to encode the logical state of the

E-mail address: bennetc@watson.ibm.com (C.H. Bennett).

1355-2198/02/$ - see front matter © 2003 Published by Elsevier Science Ltd.
doi:10.1016/S1355-2198(03)00039-X

computation, and these *information bearing degrees of freedom* (IBDF) are by design sufficiently robust that, within limits, the computer's logical (i.e., digital) state evolves deterministically as a function of its initial value, regardless of small fluctuations or variations in the environment or in the computer's other *non-information bearing degrees of freedom* (NIBDF). While a computer as a whole (including its power supply and other parts of its environment), may be viewed as a closed system obeying reversible laws of motion (Hamiltonian or, more properly for a quantum system, unitary dynamics), Landauer noted that the logical state often evolves irreversibly, with two or more distinct logical states having a single logical successor. Therefore, because Hamiltonian/unitary dynamics conserves (fine-grained) entropy, the entropy decrease of the IBDF during a logically irreversible operation must be compensated by an equal or greater entropy increase in the NIBDF and environment. This is Landauer's principle. Typically the entropy increase takes the form of energy imported into the computer, converted to heat, and dissipated into the environment, but it need not be, since entropy can be exported in other ways, for example by randomizing configurational degrees of freedom in the environment.

Landauer's principle appears straightforward, but there is some subtlety in understanding when it leads to thermodynamic irreversibility. If a logically irreversible operation like erasure is applied to random data, the operation still may be thermodynamically reversible because it represents a reversible transfer of entropy from the data to the environment, rather like the reversible transfer of entropy to the environment when a gas is compressed isothermally. But if, as is more usual in computing, the logically irreversible operation is applied to known data, the operation is thermodynamically irreversible, because the environmental entropy increase is not compensated by any decrease of entropy of the data. This wasteful situation, in which an operation that *could* have reduced the data's entropy is applied to data whose entropy is already zero, is analogous to the irreversibility that occurs when a gas is allowed to expand freely, without doing any work, then isothermally compressed back to its original volume. Fortunately, these wasteful operations can be entirely avoided: it is possible to reprogram any deterministic computation as a sequence of logically reversible steps, provided the computation is allowed to save a copy of its input. The logically reversible version of the computation, which need not use much more time or memory than the original irreversible computation it simulates, can then, at least in principle, be performed in a thermodynamically reversible fashion on appropriate hardware.

2. Objections to Landauer's principle

One of the main objections to Landauer's principle, and in my opinion the one of greatest merit, is that raised by Earman and Norton (1999, pp. 1–8), who argue that since it is not independent of the Second Law, it is either unnecessary or insufficient as an exorcism of Maxwell's Demon. I will discuss this objection further in the third section (See also Bub, 2000).

Others have argued that Landauer's principle is actually false or meaningless. These objections are of three kinds:

1. It is false or meaningless because there is no connection between thermodynamic quantities like heat and work and mathematical properties like logical reversibility, so that comparing the two is comparing apples and oranges;
2. it (or, more precisely its converse) is false because *all* data-processing operations, whether logically reversible or not, require the dissipation of at least $kT\ln 2$ of energy—and indeed usually much more—to be accomplished by any actual physical apparatus; or,
3. it is false because even logically irreversible operations can in principle be accomplished in a thermodynamically reversible fashion.

The first objection touches deep questions of the relation between mind and matter which are not entirely in the province of science, although physicists have long felt a need to address them to some extent. From its beginning, the history of the Maxwell's Demon problem has involved discussions of the role of the Demon's intelligence, and indeed, of how and whether one ought to characterize an "intelligent being" physically. On this question I will take the usual approach of physicists and banish questions about intelligence by substituting an automatically functioning mechanism whenever an intelligent being is required. Not only is this mechanism supposed to be automatic, it ought to obey accepted physical laws. In particular, the entire universe, including the Demon, should obey Hamiltonian or unitary dynamics, when regarded as a closed autonomous system. From this viewpoint, the first objection loses much of its persuasiveness, since there appears to be no deep conceptual problem in inquiring whether an automatically functioning apparatus designed to process information (i.e., a computer) can function in a thermodynamically reversible fashion, and if not, how the thermodynamic cost of operating the apparatus depends on the mathematical properties of the computation it is doing.

The second objection, that even logically reversible data-processing operations cannot be accomplished in a thermodynamically reversible fashion, I believe has largely been overcome by explicit models, proposed by myself and others, of physical mechanisms, which obey the accepted conventions of thermodynamic or mechanical thought experiments, and which accomplish reversible computation at zero cost (so-called ballistic computers, such as the Fredkin-Toffoli hard sphere model; Fredkin & Toffoli, 1982), or at a per-step cost tending to zero in the limit of slow operation (so-called Brownian computers, discussed at length in my review article; Bennett, 1982). These questions were revisited and vigorously debated in an exchange in *Physics Reviews & Letters* (Porod, Grondin, Ferry, & Porod, 1984). Of course, in practice, almost all data processing is done on macroscopic apparatus, dissipating macroscopic amounts of energy far in excess of what would be required by Landauer's principle. Nevertheless, some stages of biomolecular information processing, such as transcription of DNA to RNA, appear to be accomplished by chemical reactions that are reversible not only in principle but in practice.

Among the most important logically reversible operations are copying unknown data onto a blank (i.e., initially zero) register, and the reverse process, namely erasure of one of two copies of data known to be identical. To show that these operations are logically reversible, we write them in standard programming notation, where y represents the register to be changed and x a reference register which is either added to it or subtracted from it

$$y := y + x,$$
$$y := y - x.$$

The first operation copies the value of x into y, if y is initially zero; the second erases y to zero, if y and x are known to be initially equal. For any initial values of x and y, the two operations are logically reversible since each exactly undoes the effect of the other. Physically reversible means of performing these and other logically reversible operations are discussed in Landauer (1961) and Bennett (1982). An important example of a logically reversible operation, in the context of Maxwell's Demon and Szilard's engine (Bennet, 1982, Fig. 12), is reversible measurement, in particular the reversible transition of a memory element from a standard initial state S into one of two states, call them L and R, depending on whether the single molecule in Szilard's engine is located on the left or right. The physical reversibility of this operation in terms of phase space volumes is illustrated in Fig. 12 of Bennett (1982); an explicit clockwork mechanism for carrying out this reversible measurement is analyzed in Bennett (1987).

For the remainder of this section, I will focus on the third objection, the argument that even logically irreversible operations can be accomplished in a thermodynamically reversible fashion. This position has been asserted in various forms by Earman and Norton (1999, pp. 16–18) and Shenker (2000). In the context of a modified Szilard engine, similar to Fig. 12 of Bennett (1982), Earman and Norton consider a demon with memory capable of two states, L and R, which follows a program of steps similar to a computer program, executing one of two separate logically reversible subprograms, "program-L" and "program-R," depending on whether the molecule is found on the left or right side of the partition. Initially, they assume the memory is in state L (their state L does double duty as the standard initial memory state, called S in Bennett (1982) and as the left post-measurement state, called L in Bennett (1982)). Instructions are labeled below according to the part of the Demon's cycle. Steps M1–M4 comprise the measurement segment, R1–R5 the "right" subroutine, and L1–L4 the "left" subroutine. The memory state at the end of the each operation is shown in brackets.

M1. Insert partition [L]
M2. Observe the particle's chamber[L] or [R]
M3. If memory bit = R, go to R1 [R]
M4. If memory bit = L, go to L1 [L]
R1. Attach pulleys so right chambercan expand [R]
R2. Expand, doing isothermal work W [R]
R3. Remove pulleys [R]

R4. Transform known memory bit from R to L [L]
R5. Go to M1 [L]
L1. Attach pulleys so left chamber can expand [L]
L2. Expand, doing isothermal work W [L]
L3. Remove pulleys [L]
L4. Go to M1 [L]

At the culmination of step R4 or L3, Earman and Norton argue that both the gas and demon are back in their initial states. Work W has been done by removing energy heat from the reservoir without any other change in the universe. If this were really so, the Second Law would have been violated.

I would argue, on the contrary, that while each of the routines L and R by itself is logically reversible, the combination is not, because it includes a merging in the flow of control, which is just as much a case of logical irreversibility as the explicit erasure of data. The instruction M1 has two predecessors (L4 and R5). Therefore, when executing this program, there is a two-to-one mapping of the logical state as control passes from L4 or R5 to M1. This is where the work extracted by the demon must be paid back, according to Landauer's principle.

The fact that a merging of the flow of control constitutes logical irreversibility is also illustrated in Fig. 1 of Bennett (1982), where the final Turing machine configuration has two predecessors not because of an explicit erasure of data, but because the head could have arrived at its final location by two different shifts, one of which corresponds to the intended computation path, while the other corresponds to an allowed transition to the final state from an extraneous predecessor configuration which is not part of the intended computation.

A similar objection to Landauer's principle, this time illustrated with an explicit gear mechanism involving a pinion operating between two rigidly connected but opposing half-racks, is presented by Shenker (2000) in her Fig. 5, which is adapted from a diagram of Popper (see Leff & Rex, 1990, Chap. 1, Fig. 12). The accompanying text argues that a single external manipulation, namely a counter-clockwise rotation applied to pinion gear B, would restore a memory element called the "key" from either of two initial positions labeled, respectively, "R" and "L," to a neutral final position labeled "?," corresponding to the standard state "S" in Bennett (1987).

While it is indeed true that the counterclockwise rotation of the pinion would do this, the act of performing that rotation is not thermodynamically reversible, as one can see by considering in more detail the possible motions of and mutual constraints between the two relevant mechanical degrees of freedom, namely: (1) the rotation angle of the pinion and (2) the lateral (left-right) displacement of the key with its two rigidly attached half-racks, termed "grooves for gear B" in the figure. For any rotation angle of the pinion, there will be a finite range of backlash within which the key can rattle back and forth before being stopped by collision with the gear teeth on the pinion (to demand zero backlash would entail an infinite negative configurational entropy; the argument given here in support of Landauer's principle is independent of the amount of backlash, requiring only that it be non-zero). Consider the resetting

action when the key is initially in the "L" position. The accompanying figure shows schematically how the range of motion of the information bearing coordinate (in this case the left/right displacement of the key) varies as a function of a controlling coordinate (in this case the rotation angle of the pinion B) whose steady increase brings about a merger of two logically distinct paths of the information-processing apparatus (in this case the resetting of the key to the neutral position). At stage 1, the information bearing coordinate is confined to one of the two about to be merged paths. At stage 2, the barrier separating these paths disappears, and the information bearing coordinate suddenly gains access to twice as large a range as it had before. This is an irreversible entropy increase of $k \ln 2$ analogous to the $Nk \ln 2$ entropy increase when a gas of N atoms leaks by free expansion into twice its original volume, without doing any work. At stage 3, the controlling coordinate (pinion) does work on the information-bearing coordinate (key), eventually, by stage 4, compressing it back to its original range of motion.

Although Landauer's principle appears simple to the point of triviality, there is, as noted in the first section, a somewhat subtle distinction in how it applies depending on the nature of the initial state. As noted in Bennett (1982), Fig. 16, a logically irreversible operation, such as the erasure of a bit or the merging of two paths, may be thermodynamically reversible or not depending on the data to which it is applied. If it is applied to random data—a bit that is initially equi-probably distributed between 0 and 1, or a key that is equi-probably on the R or the L path in the accompanying figure, it is thermodynamically reversible, because it decreases the entropy of the data while increasing the entropy of the environment by the same amount. In terms of usual thermodynamic thought experiments, it is analogous to isothermal compression, which decreases the entropy of a gas while increasing the entropy of the environment. This situation arises in the application of Landauer's principle to Szilard's engine: the data being erased (R or L) is random; therefore, its erasure represents a reversible entropy transfer to the environment, compensating an earlier entropy transfer *from* the environment during the isothermal expansion phase, and making the total work yield of the cycle zero, in obedience to the Second Law. However, as Landauer and Shenker note, the data in the course of the usual deterministic digital computation is not random, but on the contrary determined by the computer's initial state.

Thus, at least in the context of knowledge of the initial state, whenever a logically irreversible step occurs in the course of a deterministic computation, one of the predecessors is certain to have been on the computation path, while the other(s) have zero chance of having been on the path. Shenker (2000) argues (Fig. 3) that the 1:1 sequence of states actually visited in the course of a deterministic computation means that deterministic computers are not bound by Landauer's principle. In fact, by the above argument, the performance of a 1:1 state mapping by a manipulation that *could have* performed a 2:1 mapping is thermodynamically irreversible, the irreversibility being associated with the wasteful instant (stage 2 in above) at which a constrained degree of freedom is allowed to escape from its constraint without any work being exacted in exchange. In fact, the significance of Shenker's and Landauer's observation (viz., that the states actually visited in a deterministic

computation comprise an unbranched chain) is somewhat different. It means that it is always possible to globally reprogram any computation that saves a copy of its input as a sequence of logically reversible steps, and therefore to perform it in a thermodynamically reversible fashion, even though the original computation, before this reprogramming, would not have been thermodynamically reversible. Reversible programming techniques, and physical models which can execute reversibly programmed computations in a thermodynamically reversible fashion, are reviewed in Bennett (1982), Bennett (1989), Levine and Sherman (1990), and elsewhere.

Another critic of Landauer's principle, Thomas Schneider (1994). argues that the energy cost of biological information processing ought not to be regarded as coming from resetting or erasure, but rather from a two step process in which a molecular computing system is first "primed" or activated by the addition of energy, and later dissipates this energy by falling into one of several final states. I would argue, on the one hand, that this is not inconsistent with Landauer's principle, and, on the other hand, that not all biochemical information-processing systems are best viewed in this way; for example, the transcription from DNA to RNA, after strand initiation, can be viewed as a logically reversible copying process driven chiefly not by prior activation of the reactant, but by removal of one of the reaction products, pyrophosphate.

Landauer's principle applies in different ways to the several kinds of physical systems capable of reversible computation or molecular scale thermodynamic engines. Such devices are broadly of three types:

(1) Ballistic computers: conservative dynamical systems like Fredkin's billiard ball computer, which follows a mechanical trajectory isomorphic to the desired computation. Such systems are incapable of merging of trajectories, so they can only be programmed to do logically reversible computations, and indeed do so at constant velocity without dissipating any energy. These devices must be isolated from external heat baths or noise sources, so they are not directly relevant to the Maxwell's demon problem.
(2) Externally clocked Brownian machines, in which a control parameter is quasi-statically varied by a macroscopic external agency to drive the system through its sequence or cycle of operations. All other coordinates are free to move randomly and equilibrate themselves within the constraint set by the value of the control parameter, and are typically coupled to a thermal bath. Many realizations of Szilard's engine, including those discussed by Earman and Norton, and that depicted in Fig. 1 of the present paper, are of this type, as are most of the proposed realizations of quantum computers. The thermodynamic cost of operating such a machine is the work done by the external agency, integrated over the cycle or sequence of operations. Mergings of trajectories are possible, but are thermodynamically costly if applied to non-random data, by the arguments given above. If such mergings are avoided through reversible programming, these devices operate reversibly in the usual sense of thermodynamic thought experiments; i.e., their dissipation per step is proportional to the driving force, tending to zero in the limit of zero speed of operation.

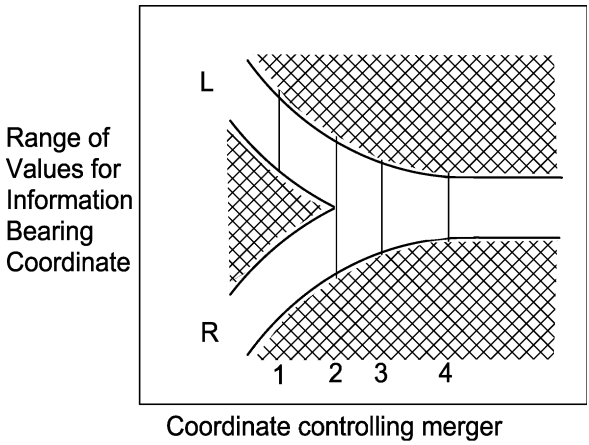

Fig. 1. Merging of two computation paths.

(3) Fully Brownian machines, in which all coordinates are allowed to drift freely, within the constraints they mutually exert on one other. This kind of machine is like an externally clocked Brownian machine where the external agency has "let go of the handle," and attached a weak spring, so the control coordinate can drift randomly forward or backward. Examples include Gabor's engine (Gabor's 1964), Feynman's classic ratchet and pawl (1963), Bennett's enzymatic and clockwork computers (1982, Figs. 6–9), and enzymes such as RNA polymerase and polynucleotide phosphorylase. The thermodynamic cost of operating this type of machine is determined by the weakest spring sufficient to achieve a net positive drift velocity. Because this type of device can drift backward and explore the tree of logical predecessors of states on its intended computation path, a small amount of merging is less costly than it would be for an externally clocked Brownian device; as described in more detail in (Bennett, 1982, Fig. 10) and at the end of Bennett (1973), such a device will drift forward, although perhaps very slowly, if driven strongly enough to prevent infinite backward excursions.

3. Landauer's principle in the context of other ideas in nineteenth and twentieth century physics

Earman and Norton have pointed out with some justice that Landauer's principle appears both unnecessary and insufficient as an exorcism Maxwell's Demon because if the Demon is a thermodynamic system already governed by the Second Law, no further supposition about information and entropy is needed to save the Second Law. On the other hand, if the Demon is not assumed to obey the Second Law, no supposition about the entropy cost of information processing can save the Second Law from the Demon. I would nevertheless argue that Landauer's principle serves an

important pedagogic purpose of helping students avoid a misconception that many people fell into during the twentieth century, including giants like von Neumann, Gabor, and Brillouin and even, perhaps, Szilard.[1] This is the informal belief that there is an intrinsic cost of order kT for every elementary act of information processing (e.g. the acquisition of information by measurement) or the copying of information from one storage medium into another, or the execution of a logical operation by a computer, regardless of the act's logical reversibility or irreversibility. In particular, the great success of the quantum theory of radiation in the early twentieth century led Gabor and Brillouin to seek an exorcism of the Demon based on a presumed cost of information acquisition, which in turn they attributed to the energy cost of a thermal photon, or in the case of Gabor's (1964) high-compression Szilard engine, to the cost of recreating a static radiation field localized to one end of a long cylinder, into which the molecule would wander to trigger the power stroke.

Landauer's principle, while perhaps obvious in retrospect, makes it clear that information processing and acquisition have no intrinsic, irreducible thermodynamic cost whereas the seemingly humble act of information destruction does have a cost, exactly sufficient to save the Second Law from the Demon. Thus, measurement and copying *can* be intrinsically irreversible, but only when they are conducted in such a way as to overwrite previous information. The Second Law, uniquely among physical principles, is and probably always will be in need of explanations and worked-out examples showing why microscopically reversible physical systems cannot escape it. When first told of the Demon, any normally curious person will be dissatisfied with the explanation. "It can't work because that would violate the Second Law" and will want to see exactly *why* it can't work. That is the virtue of worked-out examples such as Feynman's ratchet and pawl. Indeed Feynman's ratchet and pawl argument provides a more fundamental and elegant refutation of Gabor's engine than that given by Gabor, one that does not depend on the quantum theory of radiation.

Gabor's (1964), Fig. 7 engine uses an optically triggered mechanism to trap the molecule at one end of a long cylinder. Once the molecule has been trapped, it is made to do a large amount, $k_B T \ln X$, of work by isothermal expansion, where X is the expansion ratio, hopefully more than enough to replace the energy E dissipated when the trap was sprung. But no trapping mechanism, optical or otherwise, can be completely irreversible. As Feynman points out in his analysis of the fall of the pawl on the edge of a ratchet tooth, which is designed to prevent the ratchet from rotating backwards, a trapping mechanism that dissipates energy E at temperature T has a probability $\exp(-E/k_B T)$ of running backward. The cyclic operation of Gabor's engine is mathematically equivalent to a ratchet machine, which dissipates the trapping energy E when the pawl falls, and does $k_B T \ln X$ work as the ratchet rotates

[1] Szilard's classic 1929 paper is tantalizingly ambiguous in this respect. While most of the paper seems to attribute an irreducible thermodynamic cost to information acquisition—and indeed, this is how his paper has been subsequently interpreted—his detailed mathematical analysis (Leff & Rex, 1990, p. 131) shows the entropy increase as occurring during the resetting step, in accordance with Landauer's principle. Probably Szilard thought it less important to associate the entropy increase with a specific stage of the cycle than to show that it must occur somewhere during the cycle.

in the intended forward direction to the next tooth. If $E > k_B T \ln X$, Gabor's engine will run in the intended forward direction, but will not do enough work to replace the energy lost in springing the trap. If $E < k_B T \ln X$, the engine will run in the reverse of its intended direction, alternately compressing the molecule and letting it escape into the long cylinder through backward operation of the trap, and again it will not violate the Second Law.

Acknowledgements

I acknowledge support from the US Army Research office, grant DAAG55-98-C-0041 and DAAG55-98-1-0366.

References

Bennett, C. H. (1973). Logical reversibility of computation. *IBM Journal of Research and Development*, *17*, 525–532.

Bennett, C.H. (1982). The thermodynamics of computation—a review. *International Journal of Theoretical Physics 21*, 905–940. http://www.research.ibm.com/people/b/bennetc /bennettc1982666c3d53.pdf

Bennett, C. H. (1987). Demons, engines, and the second law. *Scientific American*, *257*, 108–117.

Bennett, C. H. (1989). Time/space trade-offs for reversible computation. *SIAM Journal of Computing*, *18*, 766–776.

Bub, J. (2002). Maxwell's Demon and the thermodynamics of computation. arXiv:quant-ph/0203017.

Earman, J., & Norton, J. D. (1999). Exorcist XIV: The wrath of Maxwell's Demon. Part II. From Szilard to Landauer and beyond. *Studies in the History and Philosophy of Modern Physics*, *30*, 1–40.

Feynman, R. (1963). Ratchet and pawl. *In The Feinman lectures on physics. Reading, MA: Addison Wesley*.

Fredkin, E., & Toffoli, T. (1982). Conservative logic. *International Journal of Theoretical Physics*, *21*, 219.

Gabor, D. (1964). Light and information. *Progress in Optics*, *1*, 111–153.

Landauer, R. (1961). Dissipation and heat generation in the computing process. *IBM Journal of Research and Development*, *5*, 183–191.

Leff, H. S., & Rex, A. F. (1990). *Maxwell's Demon: Entropy, Information, Computing*. Princeton: Princeton University Press.

Levine, R. Y., & Sherman, A. T. (1990). A note on Bennett's time-space trade-off for reversible computation. *SIAM Journal of Computing*, *19*, 673–677.

Porod, W., Grondin, R. O., Ferry, D. K., & Porod, G. (1984). Dissipation in computation. *Physics Reviews and Letters*, *52*, 232–235 and ensuing discussion.

Schneider, T. (1994). *Nanotechnology*, *5*, 1–18.

Shenker, O.R. (2000). Logic and entropy. http://philsci-archive.pitt.edu/documents /disk0/00/00/01/15/index.html