Craig S. Lent · Alexei O. Orlov
Wolfgang Porod · Gregory L. Snider
Editors

Energy Limits in Computation

A Review of Landauer's Principle, Theory and Experiments

Springer

Energy Limits in Computation

Craig S. Lent • Alexei O. Orlov • Wolfgang Porod
Gregory L. Snider
Editors

Energy Limits in Computation

A Review of Landauer's Principle, Theory and Experiments

Springer

Editors
Craig S. Lent
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN, USA

Alexei O. Orlov
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN, USA

Wolfgang Porod
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN, USA

Gregory L. Snider
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN, USA

ISBN 978-3-319-93457-0 ISBN 978-3-319-93458-7 (eBook)
https://doi.org/10.1007/978-3-319-93458-7

Library of Congress Control Number: 2018948357

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This book is dedicated to the memory of Rolf Landauer.

Preface

Power dissipation in microprocessors has become the dominant obstacle to the development of computation. A laptop computer can burn your lap because the power dissipated (corresponding to the heat generated) per unit area by modern microprocessors can be more than that of an electric range-top unit. To combat dissipation and keep the power density at manageable levels is now the greatest challenge facing the electronics industry today. Today, computing performance must be traded against power dissipation, so that the measure of a computing system is no longer based on speed, but on speed for a given energy input, and hence energy dissipation. Therefore, an important question is whether there is a minimum energy that must be dissipated to heat in a computational operation. The existence of such a minimum would suggest that there are fundamental limits to progress in computation. The underlying question is therefore one of the relationships between energy and information.

Discussions of the link between information and energy have a long history, going back to Maxwell's demon, proposed in the nineteenth century as a challenge to the second law of thermodynamics. Maxwell's demon suggests that by making a measurement, and then using the result of that measurement to perform a set of reversible processes, the entropy of a system can be lowered without doing net work. This is a violation of the second law, and discussions since then have tried to unravel the issue of where the demon fails, so that the second law is preserved. In 1929, Szilard proposed that when the demon does a measurement, some dissipation must occur. In the 1960s Rolf Landauer proposed that dissipation to heat must occur not as a result of measurement, but only if information is erased (destroyed), and that if information is destroyed an energy of at least $k_B T \ln 2$ must be dissipated, where k_B is the Boltzmann constant, T is the temperature, and ln is the natural logarithm. This proposal has come to be known as Landauer's principle. The energy mentioned in Landauer's principle, about 3×10^{-21} Joules at room temperature, is a very small amount of energy, negligible compared to the energy dissipated in real computers, so for decades discussions of Landauer's principle were largely academic, and always theoretical. From the beginning, Landauer's principle has been controversial, and many academic papers have been published discussing the

validity of the principle. In his original paper, Landauer makes reference to entropy in his derivations, suggesting a link between thermodynamics and information. Much of the subsequent discussion of Landauer's principle has revolved around this possible link and its validity.

While the Landauer's principle has long been a topic of academic debate, it has gained new prominence recently due to the large amount of heat generated by today's computers. If Landauer's principle is correct, there may be ways to build computers that dissipate far less than today's computers. If Landauer's principle is incorrect, there is a lower limit of heat that must be dissipated by a computer at each step, which sets an unavoidable bound on how much heat a computer must produce. To prove the validity of Landauer's principle requires real-world tests, but only in the last few years have experiments been done at energy levels that can provide a valid test.

The answer to the question of the validity of Landauer's principle has enormous implications for future research directions in electronic devices and computers, as well as for the electronics and computer industries. This book brings together all sides of the discussions regarding Landauer's principle and examines both theoretical and experimental issues. Its six chapters are authored by leaders in the discussions of energy use in computation and the physical underpinnings of information.

The first four chapters are devoted to a discussion of the link between information and thermodynamics. Chapter 1 starts with information theory and argues that thermodynamics can be constructed as a specific application of information theoretic entropy to equilibrium physical systems, quantum and classical, and supports Landauer's principle. Chapter 2 offers a proof of Landauer's principle from the unitary quantum evolution of a physical system and the thermal environment. Chapter 3 approaches the question from the point of view of the second law of thermodynamics and highlights different notions of reversibility, including the role of feedback, measurement, and control. Chapter 4 argues that, in fact, the connection between information theoretic entropy and thermodynamic entropy is illusory. On this account Landauer's principle is false—a result of using the same word in two quite different contexts. The last two chapters are primarily experimental. Chapter 5, by Ciliberto and Lutz, which includes both experiment and theory, presents a recent test of the Landauer's principle, along with experiments involving the links between information and energy. Chapter 6, by Orlov et al., presents another experimental test of the Landauer's principle and explores adiabatic reversible computing systems that avoid the destruction of information.

Discussions of the relationship between energy and information, going on for over 140 years, are more important now than ever. Computing systems, and the demand for computation, have reached levels where the energy cost of information can no longer be ignored as a mere academic issue. This book brings together a rich discussion of the foundations of the link between energy and information, and the implications this link has on the future development of computing systems.

We would like to thank all the authors who contributed chapters to this book. Without their time, effort, and insight, this book would not have been possible.

Notre Dame, IN, USA

Craig S. Lent
Alexei O. Orlov
Wolfgang Porod
Gregory L. Snider

Contents

About the Editors

Craig S. Lent is the Frank M. Freimann Chair Professor of Engineering and Concurrent Professor of Physics at the University of Notre Dame. He received his PhD in solid state physics in 1984 from the University of Minnesota. His field of research is quantum devices and molecular-scale devices. The current research of his group focuses on the fundamental theoretical limits imposed by physics on computing devices. For the past several years his group has been investigating these questions in the context of a new transistor-less paradigm known as quantum-dot cellular automata (QCA), developed at Notre Dame and now the subject of research worldwide.

Alexei O. Orlov received his M.S. in physics from the Moscow State University, Moscow, Russia, in 1983. He is currently a Research Professor at the University of Notre Dame, Notre Dame, IN, USA. From 1983 to 1993, he worked at the Institute of Radio Engineering and Electronics, Russian Academy of Sciences, Moscow, Russia. He received his Ph.D. in physics of semiconductors and dielectrics from the same institute in 1990. During that time, he conducted research on mesoscopic and quantum ballistic effects in electron transport of GaAs field-effect transistors. He was a visiting fellow at the University of Exeter, UK in 1993, and joined the Department of Electrical Engineering, University of Notre Dame, in 1994. His research interests include experimental studies of mesoscopic, single-electron and molecular electronic devices and sensors, nanomagnetics, quantum-dot cellular automata, and nanothermoelectrics. He has authored or coauthored more than 150 publications.

Wolfgang Porod currently is Frank M. Freimann Professor of Electrical Engineering at the University of Notre Dame. He received his Diplom (M.S.) and Ph.D. from the University of Graz, Austria, in 1979 and 1981, respectively. After appointments as a postdoctoral fellow at Colorado State University and as a senior research analyst at Arizona State University, he joined the University of Notre Dame in 1986 as an associate professor. He also has served as the founding director of Notre Dame's Center for Nano Science and Technology (ND*nano*). His research

interests are in the area of nanoelectronics and nanomagnetics, with an emphasis on new circuit concepts for novel devices. He has authored some 600 publications and presentations. He is a fellow of the IEEE and he has served (2002–2003) as the Vice President for Publications on the IEEE Nanotechnology Council. Over the years, he has been active in organizing conferences, special sessions, and tutorials, and as a speaker in IEEE Distinguished Lecturer Programs.

Gregory L. Snider received his Ph.D. at the University of California, Santa Barbara, in 1991 and was a post-doc at Cornell University. He joined the University of Notre Dame in 1994. His research focuses on the design, fabrication, and measurements of micro- and nanoelectronic devices and circuits. In the micro regime, his group works on CMOS circuits to study dissipation in computation, as well as to interface CMOS and nano devices. On the nano side, his research focuses on single-electron devices including quantum-dot cellular automata (QCA) and molecular devices. He is author or coauthor of approximately 300 publications and conference presentations. He has served as the associate chair for Graduate Studies in the Department of Electrical Engineering, an associate editor of the *IEEE Transactions on Electronic Devices*, and is a fellow of the IEEE.

Information and Entropy in Physical Systems

Craig S. Lent

Contents

C. S. Lent (✉)
Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA
e-mail: lent@nd.edu

C. S. Lent et al. (eds.), *Energy Limits in Computation*,
https://doi.org/10.1007/978-3-319-93458-7_1

1 Introduction: What Is Information?

When we look out into the physical world, we do not see information. Rather, we see the physical world and the way things are. Or at least that is what we can hope to see, discern, reconstruct, or model, from evidence and experimentation. According to our best theories, what we find in the physical world are particles and fields, or perhaps more correctly: fields and the particles that are the quanta of those fields.

One use of the term "information" refers to this *raw information*—the state of the physical world or part of it. Physics is naturally particularly concerned with raw information. The second and actually more common use of the term "information" denotes "*encoded information*." This is information that supervenes on the raw information and can be expressed in a set of symbols, or most fundamentally, in bits. Encoded information is the domain of information theory. Connecting the two, representing encoded information and information processing in physical systems, is often the work of applied physics and electrical engineering. The questions addressed here have principally to do with the nature of the application of information theory to physical systems and the consequences of the physical law for information processing procedures.

1.1 Raw Information

The particles and fields of the physical world may exist in a number of states, permitted by the physical law, and the particular state of a specific physical system can be specified by a set of numbers. An electron can have spin up $(+1)$ or spin down (-1) relative to a given magnetic field. Perhaps this particular rock was found on the lunar surface at this particular lunar latitude and longitude with this mass and composition. The x-component of electric field at a particular point in space has a specific value this time. The electron may be found in the left well or the right well, etc. The values which describe the state of a physical system are the *raw information* contained in the physical system itself. The raw information is often quite a lot of information. It might include, for example, the position, electron configuration, and nuclear state of each and every atom (or subatomic particle) in a piece of material.

Separability In order to discuss the raw information present in a particular physical system, it is necessary to conceptually separate the system from the rest of the physical world. The simplest case of such a separation is one in which the system is in fact completely isolated, with no physical coupling or entanglement to any other system. More commonly, we rely on an approximate separation, wherein the interaction with the environment may be minimal or at least can be reasonably well characterized. In many cases, for example, the optical field couples the system to the environment. Some of the details of the raw information about a rock on the moon is flowing out into space as photons. The rock is also in thermal contact with the

lunar surface, so the details of the motion of its individual atoms are being affected by thermal fluctuations from the underlying material.

When a system is not perfectly isolated, like the moon rock, raw information can flow to or from the environment. Some information is lost and other information is gained. Micro-bombardment has perhaps altered the moon rock so that some chemical information about its earlier constitution is no longer available. That information may have moved out into the environment carried by the raw information in photons and surface vibrations, for example. Moreover, information present in the rock now perhaps includes historical information about the environment recorded through the interactions with the environment over millennia. To the trained lunar geologist, the fine structure still visible in the sample might preserve a record of a previous cataclysm that occurred 4 billion years ago. So some raw information about earlier events in the solar system has been transferred to and stored in the rock. The information in the rock may exclude a vast number of historical scenarios. The mere existence of the moon rock means that many conceivable historical sequences, such as the sun exploding or a Mars-size planet colliding with the moon, did *not* occur.

Quantum mechanics makes separability of physical systems even more challenging—we can only point to some of the issues here. Even an isolated system will generally be in a quantum superposition state. Strictly speaking, such a system has no values of dynamical variables like position, momentum, spin, or energy, until one of these is measured. An isolated need not be in an energy eigenstate, for example—in which case it is inaccurate to say that it "has" a particular value of energy. Moreover a physical system can be quantum mechanically entangled with the some or many parts of the environment. No complete description of the quantum state of one part of an entangled system can be given. Some quantum information is shared with other subsystems and is not local to any.

Is Information Conserved? If the physical system is not completely isolated, then information is clearly not conserved in the system itself. As discussed above, information can flow into or out from the system and so we cannot from the current information reconstruct the past state of the system. It may be possible then for two different past system states to evolve into one present state, for example two different levels of excitation might relax to a single ground state. Many-to-one dynamics are possible because of the environment which can, in this case for example, absorb the excitation energy and with it information about the prior state. Of course it may be that *enough* raw information is retained that a partial reconstruction is possible. When we use a physical system as a memory device, it is a requirement that some important aspects of the past can be inferred from the current state, e.g., what was the bit most recently recorded?

In classical physics, if we imagine a complete description of an entirely isolated system, or of the whole universe conceived as an isolated system, then raw information is indeed conserved by the physical law. The classical mechanical worldview of the world as reversible machinery in motion was famously expressed by Laplace describing an "intellect" subsequently known as *Laplace's Demon*:

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.[1]

Because the microscopic laws of classical physics are reversible, we can solve the equations of motion forward or backward in time. In this sense, for an isolated system the raw information is conserved. No new raw information is generated internal to the system, and in virtue of being isolated, no raw information flows in or out. For example, an isolated container of classical gas molecules has a current state consisting of the positions and momenta of all the molecules. From this raw information about the present, the past positions and momenta can be inferred by solving the equations of motion backward in time.

Quantum mechanically for a fully isolated system, information is conserved by unitary evolution of the quantum state vector and this is time-reversible. One important caveat is that measurement of any quantity (which would presumably require interaction with another system) breaks the isolation and thus destroys the reversibility. Yet, measurements seem to happen all the time independent of humans, though we do not understand in detail what is required to produce a measurement event rather than just entanglement of the target system with the measurement system.[2] Measurement, which can be triggered by small environmental interactions, forces a quantum system to choose a new state—an eigenstate of the operator that corresponds to the measured quantity. New raw information, in the form of measurement outcomes, is created, and old quantum information is destroyed.

1.2 Encoded Information

By the term "information" we most often mean what we refer to here as "*encoded information*." Consider a clay tablet on which someone has impressed arrow-shaped indentations in different orientations, or a row of capacitors each of which holds either zero charge or $+Q$, or a street sign on which has been painted the word "Stop" or "Slow." The raw information consists of the precise shape of the tablet with its indentations, the presence or absence of electrons on each capacitor, the configuration of paint pigment on the sign. The encoded information is also present, but not as additional raw information. Encoded information supervenes on the physical, raw information, through another element—the *encoding scheme*.

[1] Pierre Simon Laplace, A Philosophical Essay on Probabilities, 1814.

[2] This is the famous Measurement Problem in quantum mechanics. The term is immediately misleading because prior to the measurement, a quantum system does not in general have an underlying value of the measured result.

An encoding scheme consists of a partition of the possible states of the physical system, the raw information, and an association between each element of the partition and abstract symbols. A particular arrangement of paint pigments is associated with the symbol "S". The partition is broad enough to include variations in the precise shape of the pigment. The binary "1" might be associated with a certain amount of positive charge stored on the capacitor, give or take a margin of error. Some regions of the systems state space have no associated symbol—the pigment is in an indiscernible pattern, or the amount of charge is too low to be clearly significant. The usual encoding scheme partitions the space of possible raw information states into areas representing symbols in a generalized alphabet, and a broad region representing *invalid*, meaning nothing is encoded there.

Encoded information is deliberate. Encoded information is therefore observer-dependent. For the information to be accessible requires access to both (a) the physical system containing the raw information, and (b) the encoding scheme to map the raw information onto a set of symbols. One or two lanterns are in the bell tower. The raw information includes their detailed construction and precise position, potentially down to the atomic level. The encoding scheme consists of the mapping "one lantern" $\rightarrow$ "The British are coming by land" and "two lanterns" $\rightarrow$ "The British are coming by sea."

If one lacks knowledge of the encoding scheme, encoded information is at least unavailable information and some would argue it is not information at all, even though the raw information is present. Prior to discovering the Rosetta stone, Egyptian hieroglyphics were just raw information—patterns on walls.

A standard disclaimer: information theory is not about the *semantic content* of a string of symbols, it is only concerned with the "size" of the container. Consider the two sentences below.

```
That man wears red hats.
All men wear black hats.
```

The second sentence conveys much more information than the first, in the colloquial sense of meaningful and consequential knowledge. But because both sentences contain the same number of symbols, they have the same information theoretic size (at least prior to any further possible compression). Neither are we concerned with the truth or falsity of the information as it is connected with the way things actually are. The second sentence is certainly false. Information theory is not concerned with categories like false information or disinformation.

Reversibility of an operation on bits (a computation) is a mathematical feature of the operation. If one can correctly infer from the output of the operation what the input symbols were, then we say the process is logically reversible. The Landauer principle connects logical reversibility (input symbols can be inferred from output symbols) to physical reversibility (the physical process can be run backwards to reconstruct the input state). If raw information is transferred from the computational system to the large and complex environment, it cannot be reconstructed and so has been irreversibly lost.

Biological Information DNA encodes information for the synthesis of proteins with an encoding scheme involving "codons" composed of three-nucleotide sequences. It is now common to describe many processes in living systems as information systems—signaling, replication, sensing, transduction, etc. Information here is usually encoded in structure (as in DNA or RNA) or through the varying concentration of specific molecules. This is, of course, just a way of speaking at a higher level about raw information in chemical reactions. We normally understand this to be information conceived by *analogy* to that deliberately encoded by humans, which is taken to be encoded information sensu stricto.

1.3 Present Strategy

Our goal here is to connect what we know about the evolution of raw information, guided by the physical law, and the encoded information that supervenes on it. The particular focus here is on the Landauer Principle that connects a logical erasure of encoded information with the physical transfer of heat to the environment. Why should that be? Both the logical process and the physical process involve the concept of entropy. But entropy was defined as a physical thermodynamic and statistical mechanical quantity by Clausius, Boltzmann, Gibbs, and von Neumann, and only later defined by Shannon as an information theoretic quantity. Some argue that information theoretic entropy has nothing to do with thermodynamic entropy so that the Landauer Principle makes a category error simply because the two words are identical [1]. Norton argues that mistaking unknown bits for a "bit gas" and thereby confusing the two entropy concepts is simply silly [2].

To disentangle this requires several steps. The next section attempts to articulate carefully the concept of probability, which has both information theoretic and physical uses. Section 3 introduces the Shannon notion of entropy, here called the Shannon measure of information (SMI) as a measure on a probability distribution. The Jaynes principle of maximum entropy is then used for the information theoretic problem of constructing a probability distribution given limited knowledge about outcomes that can be expressed as mathematical constraints. The results are of immediately familiar to anyone acquainted with statistical mechanics. Section 4 follows Jaynes path in making the connection to physics. This can then be applied to the Landauer Principle as discussed in Sect. 5. A key result here is a concrete and specific calculation of entropy and heat flow in a minimal physical system. The quantum formulation in Sect. 6 requires extension of the basic formalism to open systems. The connection to Shannon entropy is made both through the usual von Neumann entropy and through the less-familiar "entropy of outcomes." The quantum calculation of free expansion of a gas is revealing in this regard. By grounding statistical mechanics explicitly in the information theoretic notion of entropy, we can firmly establish the connections that make the Landauer Principle clear and compelling.

2 Probability

We first consider two classical systems.

System 1: A Fair Die A fair six-sided die is randomly cast on a table top. The possible results are $1, 2, \ldots, 6$, and the probability of obtaining each result is identical.

$$P_1 = P_2 = P_3 = P_4 = P_5 = P_6 = 1/6 \tag{1}$$

System 2: An Ideal Gas We consider a monatomic gas with a very large number N of non-interacting atoms (e.g., argon) in a volume V with pressure P. Let us assume the system is in thermal contact with a heat bath with temperature T. A heat bath is a system with thermodynamically many degrees of freedom that has long-since stabilized all average measures. If the accessible microstates of the system are enumerated and have energies E_i, then the well-known Boltzmann result (to which we will return) is that the probability of finding the system in state j is

$$P_j = \frac{e^{-E_j/k_B T}}{\sum_j e^{-E_j/k_B T}}. \tag{2}$$

It is worth noting a few features of this basic description. We take it as understood in the classical case that at any particular time a specific system indeed *has* a specific state, and that the state which follows is determined by the previous state and the physical law. At the moment I toss the die into the air it has a certain position, velocity, and angular momentum about some axis. Knowing that, as well as perhaps the details of the air currents in the room and the landing surface properties, one could imagine calculating the precise trajectory including any bouncing on the table. The resulting motion, right through until the die settles on the surface, could in principle be calculated, and was certainly determined at the moment the die left my hand.

Similarly, for the ideal gas: the positions and momenta of all the N particles constitute the actual, objective, state of the system at a particular time. There is a "fact of the matter" as to what the microstate of the gas (this liter of argon on my desk) is right now. It is a practical impossibility for us to measure all these dynamical quantities, particularly at the same instant, but they presumably exist.

We use the language and calculus of probabilities because we lack a complete knowledge of the state and its change over time. The probabilities are therefore "observer-relative." A robotic die tosser with fairly precise launching techniques might be able to predict, say, that the outcomes would more likely be a $4, 5$, or 6. An observer who knew more microscopic information about the ideal gas could reasonably assign a different set of probabilities to the microstates of the system. Equation (2) represents the probabilities that *any* observer who knew only the macroscopic quantities T, N, and V should assign to the microstates of the specific system in front of them. It is not in that sense "subjective." It does not depend on who the observer is or on their emotional or mental state.

Laplace's demon, who knows the position and momentum of each particle in the universe, has no need of probabilities. The physical law as understood in classical mechanics enables the computation of all future states from a complete description of the present state. It is a practical impossibility to make such a calculation, given human limitations and also the limitations of the physical resources available in the universe for computation. But the point of principle is important. The classical universe is simply solving its equations of motion forward in time.

We are adopting a Bayes/Jaynes approach here that probabilities are to be understood as numerical expressions of partial knowledge, incomplete information, of a present state, or a future event. A probability $P = 1$ represents certain knowledge that the event *will* occur, a probability $P = 0$ means the event certainly *will not* occur, and a real value between 1 and 0 represents greater or lesser partial knowledge that the event will occur. Equations (1) and (2) specify the probability for a future measurements of the state of each system.

On the classical account, a measurement of the system (e.g., looking at the die) reveals an existing fact of the system's state that was true the instant before the measurement occurred. Therefore, we do not need to distinguish between the probability of a measurement event having a certain outcome and the system having a certain state. We can equally well talk about the probability of the die being on the surface with a 5 showing and the die being seen to be a 5, or revealed to be a 5 when a shaker cup is lifted. The quantum account, discussed in Sect. 6, is different.

Confirmatory evidence that a probability distribution was correct would be the relative frequencies of many such measurements on many essentially identically prepared systems. For the die of (1), that would take many tosses of a fair die. For the thermodynamic case of (2), that means with the same macroscopic variables—an ensemble average in the limit of many trials.

Another feature of this probabilistic analysis is revealed in the phrase "accessible microstate." There is always a background knowledge of the system which precedes the assignment of probabilities and limits the set of possibilities considered to what we will call the accessible region (AR). In the case of the die, for instance, we are assuming that the die will in fact land on the table and have a face up. We decide to ignore other possible sequences of events. Perhaps it falls off the table, bounces and lands tilted up against a table leg with a corner of the die facing upward. Perhaps a meteor impacts the table with catastrophic results before the die can land. For the gas, we assume of course that the container doesn't leak, that a passing ultra-high energy cosmic ray doesn't deposit extra energy in the gas, etc. A set of extremely low-probability possibilities are removed from consideration at the outset, normally without comment. The AR must be kept in mind because what constitutes "complete ignorance" about the outcome is, as in the case of the die above, uniform probability over the AR, not uniform probability over *every conceivable* outcome.[3] We always begin some background knowledge (Fig. 1).

[3] We will not wade into the subtler issues involved, but refer the reader to Chapter 12 of Jaynes [3]. The quantum treatment in Sect. 6 actually makes the choice of basis explicit, and therefore clarifies the question: "Ignorance with respect to what?"

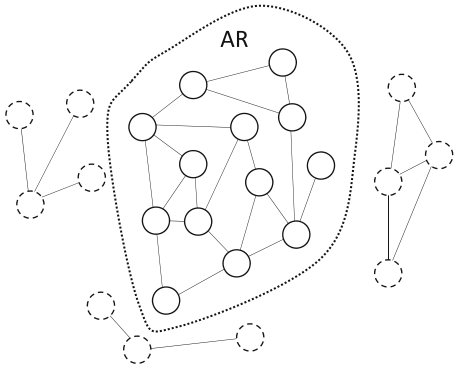

Fig. 1 Schematic of accessible region (AR) of state space. Circles represent possible system state and line represents possible transitions. Some state, those shown outside the dotted line, are reasonable to practically ignore because they are either too rare or difficult to access

3 Information Theory

3.1 SMI: The Shannon Measure of Information

Claude Shannon, considering the transmission of symbols in communication, introduced a measure on a probability distribution which he called the entropy. Using the term "entropy" was well-motivated, and was the course of action advised by von Neumann, but it has resulted in some confusion. We will adopt the strategy of Ben Naim and call this measure the Shannon Measure of Information (SMI) [4, 5].

The SMI characterizes a probability distribution $P = [P_1, P_2, \ldots, P_k, \ldots, P_N]$ by a real non-negative number, measured in bits, computed from the distribution.

$$\boxed{\text{SMI}[P] = -\sum_{k=1}^{N} P_k \log_2(P_k)} \tag{3}$$

The SMI is a measure of the amount of information, in bits, that one is missing if all one knows about the current state or future outcomes is the probability distribution P. If one outcome, say event 2, is certain, then $P_2 = 1$ and $P_k = 0$ for all other k. In that case the SMI is 0; there is no missing information. If all the probabilities are equal, $P_k = 1/N$ for all k and the SMI is $\log_2(N)$ bits. If N is an even power of 2, this is clear: $N = 4$ corresponds to 2 bits missing; $N = 8$ corresponds to 3 missing bits, etc.

Figure 2a and b shows graphically the cases of a probability distribution among eight outcomes for the case when the outcome is certain to be event 2

$$P_{(2)} = [0,\ 1,\ 0,\ 0,\ 0,\ 0] \qquad \text{SMI} = 0 \text{ bits}, \tag{4}$$

and when all outcomes are equally likely:

$$P_{\text{uniform}} = \left[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right] \qquad \text{SMI} = 3 \text{ bits}. \tag{5}$$

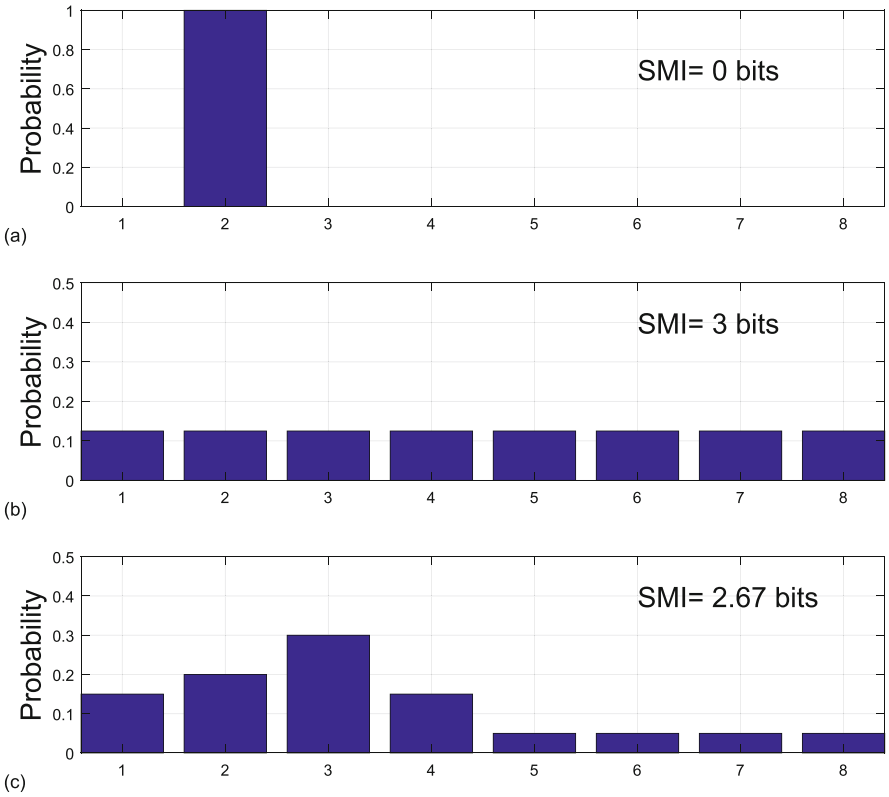

Fig. 2 Probability distributions and associated Shannon Measure of Information (SMI). (**a**) If exactly one state has unit probability, then there is certainty about which state the system will be found in and no information is missing; the SMI is 0. (**b**) The case of uniform probability for eight possible states has an SMI= $\log_2(8) = 3$. (**c**) In general the probability distribution reflects some missing information, but less than complete ignorance. The amount of missing information is quantified by the SMI

Figure 2c shows the case when the probability distribution is

$$P = [0.15,\ 0.20,\ 0.30,\ 0.15,\ 0.05,\ 0.05,\ 0.05,\ 0.05] \qquad \text{SMI} = 2.67 \text{ bits}. \quad (6)$$

The SMI is intermediate between the uniform $N = 4$, SMI= 2 and the uniform $N = 8$, SMI=3 cases. With this probability distribution we know *something* about which events are likely to occur. There is an 80% chance that the result will be events 1–4, for example. We have somewhat less *missing information* that if we only knew P_{uniform}. It is convenient that SMI could also stand for "Shannon Missing Information."

3.2 SMI and the Question Game

To understand how the phrase "missing information" can have a precise meaning, it is helpful to consider a variation of the game of 20 questions and see how the SMI functions both to play the game and to make predictions. We consider the Question Game in which a player called the *chooser* selects one of set of N numbered items and the player called the *questioner* asks a series of yes/no questions with the object of deducing the index of the item chosen with the fewest number of questions.

Suppose, for example, $N = 8$ and the chooser picks an item at random, i.e. the probability for each choice is $1/8$ as in Fig. 2b. One strategy for the questioner is to ask "Is it 1?", then "Is it 2?", then "Is it 3?", and so on. On average the questioner would ask $N/2$ questions before learning the choice. This is, of course, a poor strategy.

The optimal strategy for a uniform probability distribution is the familiar binary search using repeated bipartitions. The questioner asks "Is the item in the set $\{1, 2, 3, 4\}$?", and if the answer is yes, asks "Is it in $\{3, 4\}$?", and if the answer is no, asks "Is it item 1?", and thereby has determined the choice using only three questions. This will work every time. The SMI of the uniform probability distribution over eight choices, $\log_2(8) = 3$ bits, is the number of yes/no questions one needs to ask to determine the choice using the optimal strategy. The amount of *missing information* was initially 3 bits. With the answer to each question, the questioner received an additional 1 bit of information, until finally there was no information missing and the identity of chosen item was certain to the questioner.

Suppose the selector was not selecting entirely at random, but was making the choice according to the probability distribution of Eq. (6) shown in Fig. 2c. We could imagine that the chooser is randomly drawing numbered balls from a large container. Each ball has a number $[1, \ldots, N]$ on it but there are more 3's than 2's and so forth according to the ratios in (6). The questioner knows the probability distribution. The binary search as above is now not the optimal strategy. The set $\{1, 2, 3, 4\}$ has a total probability of 80%, so asking the first question as above seems like almost wasting a question—the answer is not providing as much information.
The optimal strategy is now as follows:

1. Let the set S be the set of possible items $\{1, 2, \ldots, N\}$ and $P_k, k = 1, 2, \ldots N$ be the probabilities that each item is selected.
2. Consider all possible bipartitions of the set S into two non-empty sets, S_{left} and S_{right}.
   For each bipartition $S \rightarrow \{S_{\text{left}}, S_{\text{right}}\}$:
   (a) Sum the probabilities of the individual events in each set and renormalize to get two numbers: P_{left} and P_{right} that sum to 1.
   (b) Calculate the SMI of the probability distribution for the bipartition $P_{\text{bp}} = [P_{\text{left}}, P_{\text{right}}]$ using Eq. (3).
3. Choose the bipartition with the largest SMI and ask the chooser the question: "Is the item in S_{left}?".

4. If the answer is yes, replace S with S_{left}.
   If the answer is no, replace S with S_{right}.
5. Repeat from step (2) until there is only one item in the set S.

For our example using (6), an initial bipartition

$$S \rightarrow \{\{1, 2, 3, 5\}, \{4, 6, 7, 8\}\}$$

has an SMI of 0.88129 but

$$S \rightarrow \{\{3, 4, 5\}, \{1, 2, 6, 7, 8\}\}$$

has an SMI of 1.0, making the corresponding question a very productive question whose answer yields a full bit of information.

Now we imagine the chooser and the questioner playing the game many times with the same probability distribution P (we may suppose the numbered balls are replaced as they are drawn). Many times the optimal bipartition of the remaining set has an SMI of less than 1. Sometimes the questioner gets lucky and is able to deduce the chosen item in 2 questions, and sometimes it takes 3 or 4. Over very many games, what is the average number of questions required? One might have hoped the answer would be SMI[P], but it is not quite that simple. In one simulation, an average over 50,000 games yields an average number of questions $\langle N_{\text{q}} \rangle = 2.70$, whereas the SMI is 2.67. The constraint is actually

$$\text{SMI}[P] \; < \; \langle N_{\text{q}} \rangle \; < \; \text{SMI}[P] + 1 \tag{7}$$

which of course becomes relatively tight for large N.

We can interpret the series of yes/no answers in the game as 1's and 0's encoded in a string stored in a binary register. (We will assume that we have compressed the string using an optimal scheme—a Huffman code—so that common question sequences are encoded with fewer bits than rare sequences.) The average number of questions is bounded by SMI[P] + 1. Since the SMI need not be integer and we can only have an integer number of bit positions in the register, we need to round up to the nearest integer. We conclude that the average size of a binary register necessary to specify a particular item choice, given probability distribution P, is

$$N_{(\text{binary register})} = \text{ceil}(\text{SMI}[P] + 1). \tag{8}$$

The SMI is a quantitative measure of missing information. It is helpful to keep in mind which party is missing the information. Who is missing the information? The chooser is holding the item in her hand. She is not missing any information, for her there is complete certainty about what the chosen item is and the SMI is 0. It is the questioner who is missing the information and relying on knowledge of the probability distribution to ask good questions. The SMI in this case is a measure of *questioner's* missing information about the item, when his information about which item is chosen is incomplete and characterized by probability distribution P.

Jaynes makes this point about Shannon's original problem of a sender transmitting a message, encoded in a set of symbols, through a communication channel to a receiver.[4] The sender knows everything about the content of the message; there is no probability distribution involved and no missing information for him. Prior to getting the message, the receiver, by contrast, knows nothing about the content of message, perhaps not even the language that the message will be in. Normally the *designer* of the communication system does not know in advance what the specific messages will be, but suppose the designer does know something about the message, for example the language of the message and the probabilities of the occurrence of each letter in that language. He can then use the SMI of that probability distribution to create an efficient encoding scheme (i.e., a data compression algorithm). The SMI of the language characterizes the incomplete information of the designer of the information system. If the designer knows more about the messages, the SMI is less and he can make an even more efficient system. The sender has complete information, the receiver has no information (yet), and the designer has partial information characterized by a probability distribution and its associated SMI.

3.3 Information Gain

For the Question Game above, the SMI characterized the initial missing information of the receiver when all he knew was the probability distribution P (Eq. (6), Fig. 2c). If the receiver had no information at all, the probability distribution he would have to use is just the uniform probability P_{uniform} of Eq. (5). We can therefore ask how much information did he gain when he was given P. We define the information gained from the knowledge of the probability distribution P as the difference between the SMI (missing information) of P_{uniform} and the SMI of P.

$$I[P] \equiv \text{SMI}[P_{\text{uniform}}] - \text{SMI}[P] \tag{9}$$

$$I[P] = \log_2(N) + \sum_{k=1}^{N} P_k \log_2(P_k) \tag{10}$$

We are here using the uniform probability distribution over the AR to represent complete ignorance and asking how much information was gained by having been given the probability distribution P. In information theory, this quantity is known as the relative entropy of P with respect to $P_{(\text{uniform})}$, or the Kullback–Leibler divergence between the two [6].

[4] Jaynes [3], p. 634.

3.4 Shannon Measure for Continuous Distributions

Can we define the SMI of a probability density $P(x)$ defined for a continuous outcome x?

A natural approach is to consider the SMI of a finite discretization of $x \in [0, L]$ at each point $x_k = k\Delta x$. We note that the probability density $P(x)$ will now have the units inverse to the units of x. The probability of the event occurring in the small interval of width Δx around x_k is $P_k = P(x_k)\Delta x$. The SMI on this discrete set of $N = (L/\Delta x) + 1$ outcomes can be written

$$\mathrm{SMI}[P^{\Delta x}(x)] = -\sum_k (P(x_k)\Delta x)\log_2[(P(x_k)\Delta x)]. \tag{11}$$

If we try to take the limit of this expression as $\Delta x \to 0$, however, this quantity diverges.

Shannon suggested, but did not derive, the following expression, usually called the *differential entropy* for a probability density P(x).

$$\mathrm{S}^{\mathrm{diff}}[P(x)] = -\int P(x)\log_2[P(x)]dx \tag{12}$$

This normally converges, but turns out to have several problems if it is interpreted as a direct analogy to the discrete SMI: (a) the units of the expression are not correct, (b) it can be negative for some distributions, (c) it is not invariant under a change of variables, and (d) it does not smoothly match the discrete case in the usual Riemann limit of (11).

The fundamental problem with formulating the amount of missing information, an SMI, for a continuous distribution is simply that a particular value of x from a continuous range takes an infinite amount of information to specify. There are an infinite number of yes/no questions required to specify an arbitrary point on the real axis. So the answer to the question *"How much information is missing if all I know is the continuous probability distribution $P(x)$?"* turns out to be an infinite number of bits.

It can be argued that for any physical system, the precision of measurement (or quantum effects) limits the distinguishable values of the measurable x to a finite discrete set, so (11) is always the relevant measure.

We can, however, clearly establish the related measure $I[P(x)]$ for a continuous distribution by analogy with (9). We take the probability density reflecting complete ignorance to be the uniform distribution on the accessible regions $x = [0, L]$ to be $P_{\mathrm{uniform}}(x) = 1/L$. We can then define the information gain of $P(x)$ for both the continuous probability density and its finite discretization.

$$I[P(x)] = \mathrm{S}^{\mathrm{diff}}[P_{\mathrm{uniform}}(x)] - \mathrm{S}^{\mathrm{diff}}[P(x)] \tag{13}$$

$$I[P^{\Delta x}(x)] = \mathrm{SMI}[P^{\Delta x}_{\mathrm{uniform}}(x)] - \mathrm{SMI}[P^{\Delta x}(x)] \tag{14}$$

Taking the difference removes the problems mentioned above and (14) is numerically equivalent to a trapezoidal integration of (13) with a discretization of Δx.

3.5 Jaynes Maximum Entropy Principle

Probability is an expression of incomplete information. Given that we have *some* information, how should we construct a probability distribution that reflects that knowledge, but is otherwise unbiased? The best general procedure, known as Jaynes Maximum Entropy Principle (better would be: Maximum SMI Principle), is to choose the probabilities p_k to maximize the SMI of the distribution, subject to constraints that express what we do know.

3.6 The Microcanonical Ensemble

The simplest case is the one in which we know nothing but the rule for probabilities—that they must add up to 1.

Let us define the Shannon measure with the natural logarithm as the base, a simple matter of multiplying by $\log(2)$ (we take $\log(x)$ to denote $\log_e(x)$). The quantity of missing information represented by S_{MI} is then measured in *nats* rather than *bits*.

$$S_{\mathrm{MI}} \equiv \log(2)\mathrm{SMI} = -\sum_k p_k \log(p_k) \tag{15}$$

We want to write a probability density $P = \{p_k\}$ that maximizes $S_{\mathrm{MI}}(P)$ subject only to the constraint:

$$\sum_k p_k = 1 \tag{16}$$

Using the method of Lagrange multipliers we construct the Lagrangian

$$\mathscr{L}(P, \lambda_0) = -\sum_k p_k \log(p_k) - (\lambda_0 - 1)\left(\sum_k p_k - 1\right) \tag{17}$$

where $1 - \lambda_0$ is the Lagrange multiplier.[5] We maximize $\mathscr{L}$ by setting the partial derivatives with respect to each of the p_k to 0. The equation $\partial\mathscr{L}/\partial\lambda_0 = 0$ just recovers Eq. (16).

$$\frac{\partial}{\partial p_k}\mathscr{L}(P, \lambda_0) = -\log(p_k) - 1 + 1 - \lambda_0 = 0 \tag{18}$$

The solution is then

$$p_k = e^{-\lambda_0} \tag{19}$$

which is true for all k, so each probability is the same and, using (16) again, we have

$$\sum_k p_k = \sum_k e^{-\lambda_0} = Ne^{-\lambda_0} \tag{20}$$

$$e^{-\lambda_0} = \frac{1}{N} \tag{21}$$

$$\lambda_0 = \log(N) \tag{22}$$

$$p_k = \frac{1}{N}. \tag{23}$$

Thus our intuition that if we know nothing about the probability distribution we should make all probabilities equal is recovered from the maximum entropy principle. The value of S_{MI} at this maximum is:

$$S_{\mathrm{MI}}^{\max} = -\sum_k \frac{1}{N}\log\left(\frac{1}{N}\right) \tag{24}$$

$$S_{\mathrm{MI}}^{\max} = \log(N) \tag{25}$$

The reader will recognize in Eq. (25) the famous Boltzmann expression for entropy ($S = k_b \log W$), without the Boltzmann constant. It also connects very simply to the case in Fig. 2b where for N results with equal probability we have an SMI= $\log_2(N) = 3$ bits, the size of the binary register needed to specify one outcome, and the average number of yes/no questions needed to determine one result.

In terms of choosing balls from an urn our picture is this. An urn contains a very large number of balls (many more than N), each of which is labeled with a number $k \in \{1, 2, \ldots N\}$. There are the same large number of balls with each index, so drawing a ball randomly from the urn picks one of possible results with equal probability.

[5] The factor $(1 - \lambda_0)$ is used instead of λ_0 to simplify the form of the result.

3.7 The Canonical Ensemble

Here we consider that each ball in the urn has written on it both the index k and a value of another quantity we will call A. The values of A are denoted $[a_1, a_2, a_3, \ldots a_N]$. Every ball with a 1 on it has a_1 written on it, and so for each of the other indices. Suppose that we know the average value of the quantity A, denoted $\langle A \rangle$. By average here we mean simply an average of the values obtained from many repeated drawing of balls from the urn. What is the optimal (maximum SMI) probability distribution p_k that will yield the given average $\langle A \rangle$?

Following the maximization procedure, we maximize S_{MI} (15) subject to the two constraints:

$$1 = \sum_k p_k \tag{26}$$

$$\langle A \rangle = \sum_k p_k a_k. \tag{27}$$

We construct the Lagrangian, which now has Lagrange multipliers λ_0 and λ_1

$$\mathscr{L} = -\sum_k p_k \log(p_k) - (\lambda_0 - 1)\left(\sum_k p_k - 1\right) - \lambda_1 \left(\sum_k p_k a_k - \langle A \rangle\right) \tag{28}$$

Maximizing $\mathscr{L}$ with respect to each p_k, we obtain:

$$\frac{\partial \mathscr{L}}{\partial p_k} = -\log(p_k) - 1 + 1 - \lambda_0 - \lambda_1 a_k = 0 \tag{29}$$

with the result

$$p_k = e^{-\lambda_0} e^{-\lambda_1 a_k}. \tag{30}$$

We define

$$Z \equiv e^{\lambda_0} \tag{31}$$

or

$$\lambda_0 = \log(Z) \tag{32}$$

We call Z the partition function. The probabilities can therefore be written

$$\boxed{p_k = \frac{e^{-\lambda_1 a_k}}{Z}.} \tag{33}$$

From the constraint (26) we require

$$\sum_k p_k = 1 = \frac{1}{Z}\sum_k e^{-\lambda_1 a_k} \tag{34}$$

so

$$\boxed{Z = \sum_k e^{-\lambda_1 a_k}} \tag{35}$$

and

$$\boxed{p_k = \frac{e^{-\lambda_1 a_k}}{\sum_k e^{-\lambda_1 a_k}}} \tag{36}$$

which is the well-known Boltzmann distribution.
If we take the logarithm of Z, we obtain a way to express the constraint that the average value of A is fixed.

$$\begin{aligned}\frac{\partial}{\partial\lambda_1}\log(Z) &= \frac{\partial}{\partial\lambda_1}\sum_k e^{-\lambda_1 a_k} \\ &= \sum_k (-a_k)e^{-\lambda_1 a_k} = -\langle A\rangle\end{aligned} \tag{37}$$

hence:

$$\boxed{\langle A\rangle = -\frac{\partial}{\partial\lambda_1}\log(Z).} \tag{38}$$

We define

$$\boxed{F \equiv -\frac{1}{\lambda_1}\log(Z).} \tag{39}$$

Now we can substitute the probability distribution (36) in the resulting probability distribution to evaluate the value S_{MI} at this maximum.

$$S_{\mathrm{MI}}^{\max} = -\sum_k p_k \log(p_k)$$

$$
\begin{aligned}
&= -\frac{1}{Z}\sum_k e^{-\lambda_1 a_k}\log\left(\frac{e^{-\lambda_1 a_k}}{Z}\right)\\
&= -\frac{1}{Z}\sum_k e^{-\lambda_1 a_k}\left[\log\left(e^{-\lambda_1 a_k}\right) - \log(Z)\right]\\
&= -\frac{1}{Z}\sum_k e^{-\lambda_1 a_k}(-\lambda_1 a_k) + \underbrace{\frac{1}{Z}\sum_k e^{-\lambda_1 a_k}}_{1}\log(Z)\\
&= \lambda_1 \frac{1}{Z}\sum_k e^{-\lambda_1 a_k} a_k + \log(Z)
\end{aligned}
$$

$$
\boxed{S_{\mathrm{MI}}^{\max} = \log(Z) + \lambda_1 \langle A\rangle} \tag{40}
$$

Notice the contrast between (40) and (25). For the case with no constraints, the maximum SMI was $\log(N)$. The term $\log(Z)$ in (40) is the sum over the N exponentials shown in (35), which rapidly decrease in magnitude. This term $\log(Z)$ has a correspondingly much smaller value than $\log(N)$; the missing information is much less. Equation (25) is recovered from (40) when $\lambda_1 = 0$ and the distribution is again uniform.

Writing Eq. (40) in terms of F we obtain

$$
\boxed{F = \langle A\rangle - \frac{1}{\lambda_1} S_{\mathrm{MI}}^{\max}}. \tag{41}
$$

As an example, take the case of $N = 8$ and let the values of A be

$$
a = [\ 0.0,\ 1.0,\ 1.3,\ 2.1,\ 2.8,\ 3.4,\ 4.0,\ 6.0\ ]. \tag{42}
$$

Suppose we know that $\langle A\rangle = 1.54$. The probability distribution p_k which maximizes S_{MI} is the exponential distribution (36) with $\lambda_1 = 0.4$. This is the probability which represents just the known facts: the probabilities must add to one and the average value of A is given. Figure 3 shows the probabilities p_k associated with each of the values of a_k.

There is a one-to-one relationship between λ_1 and $\langle A\rangle$ given by Eqs. (27) and (36) and shown in Fig. 4. As λ_1 becomes large, the probability accumulates in the lowest values of A. When $\lambda_1 = 0$ we recover the uniform distribution. A large negative λ_1 results in more probability at the higher values of A. If all that is known is $\langle A\rangle$, one must just read off the associated λ_1 from Fig. 4 and assign the probability distribution appropriately using (36). Any other probability distribution

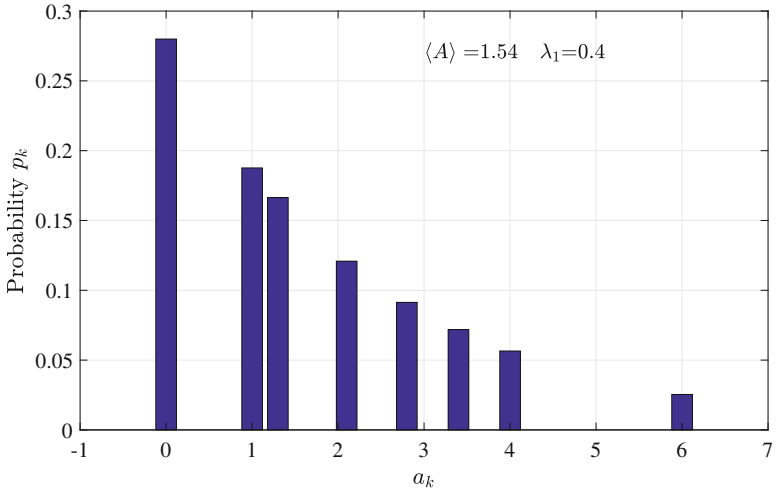

Fig. 3 The probability distribution p_k which maximizes the Shannon entropy. In this example we assume that each possible outcome of the quantity A is in the set $a = [\ 0.0,\ 1.0,\ 1.3,\ 2.1,\ 2.8,\ 3.4,\ 4.0,\ 6.0\ ]$. The distribution shown is the one that maximizes the Shannon information theoretic entropy S_{MI}, subject to the constraint that the average outcome $\langle A \rangle$ is known to be 1.54. The result is a Boltzmann distribution (36) with the Lagrange multiplier λ_1 (see (28)) equal to 0.4. This probability distribution uniquely captures only the known information. Any other distribution with the same $\langle A \rangle$ would implicitly, and incorrectly, represent more knowledge than simply knowing the average

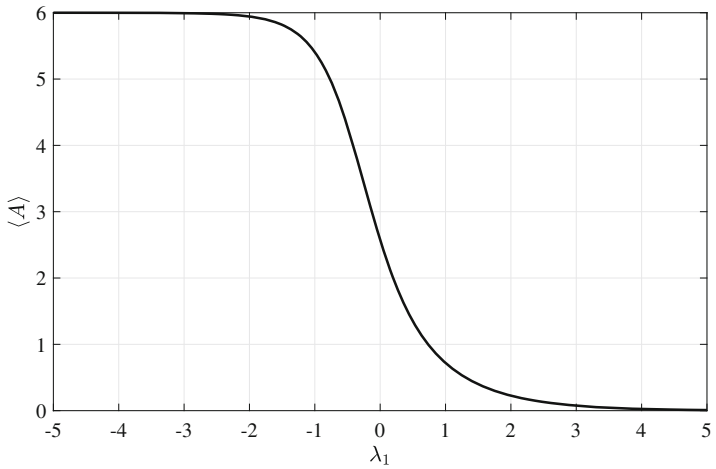

Fig. 4 For the example shown in Fig. 3, the expectation value (average) of the quantity A is shown as a function of the Lagrange multiplier λ_1. Equations (27) and (36) fix the relationship between these quantities When λ_1 is positive, the probability distribution is weighted toward small values of A. When λ_1 is negative, the probability distribution is weighted toward large values of A. When λ_1 is 0, the probability is uniform for all values of A

would implicitly assume knowledge one does not actually have—it would put in an incorrect, if unintentional, bias. (Tribus[6] calls the Lagrange multiplier λ_1 the "temper" of the distribution, an act of dramatic foreshadowing [7]).

The procedure above can straightforwardly be extended to the case when each result is labeled with the values of additional quantities that characterize the outcome. If we have another quantity B with values b_k and a known average value $\langle B \rangle$, then we would obtain the corresponding exponential distribution with an additional parameter λ_2.

$$p_k = \frac{e^{-(\lambda_1 a_k + \lambda_2 b_k)}}{\sum_k e^{-(\lambda_1 a_k + \lambda_2 b_k)}} \tag{43}$$

with

$$Z = \sum_k e^{-(\lambda_1 a_k + \lambda_2 b_k)} \tag{44}$$

and

$$\langle B \rangle = -\frac{\partial}{\partial \lambda_2} \log(Z). \tag{45}$$

$$S_{\mathrm{MI}}^{\max} = \lambda_1 \langle A \rangle + \lambda_2 \langle B \rangle + \log(Z) \tag{46}$$

The extension to any number of such quantities proceeds in the same way.

4 Classical Statistical Mechanics

We now turn to the application of the previous section to physical systems. The main results of the previous section are familiar mathematical forms from statistical mechanics. We consider now a physical system in equilibrium with a much larger system and apply the analysis to derive thermodynamic results. It will then be possible to see how to extend the analysis to metastable memory systems and dynamic systems far from equilibrium.

[6]Myron Trebus's thermodynamics text for engineers was an early attempt to popularize Jaynes grounding of the field on Shannon information theory.

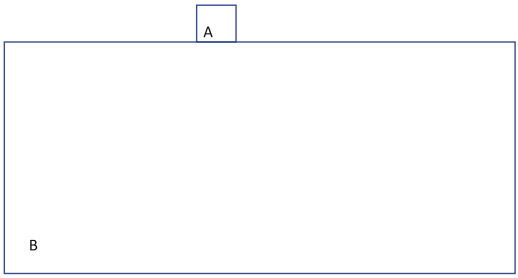

Fig. 5 A physical system A and very large system B in thermal contact so that energy can flow between them. The larger system B acts as a heat bath. In addition, Sect. 4.2 considers the case of diffusive contact, in which both energy and particles can be exchanged between the systems

4.1 Statistical Mechanics of the Canonical Ensemble

Equilibrium with a Thermal Bath Consider a physical system A in thermal contact and equilibrium with a second physical system B as shown schematically in Fig. 5. By thermal contact we mean that the two systems can exchange energy with each other. By equilibrium we mean that whatever transients occurred when they were put in contact are now over and the expectation value of all physical quantities are now time-independent. That this happens is based on our empirical experience of the physical world. System B will represent a thermal bath with very many (e.g., 10^{23}) degrees of freedom.

Suppose A can exist in N states with energies $[E_1, E_2, \ldots E_k, \ldots, E_N]$ and similarly for the bath B. We allow that different states k and j may have the same energy. The energy E_k of each state plays the role of the label on each ball a_k in urn described in the previous section. The energy of neither A nor B is fixed because energy can fluctuate between them. Because of the fluctuations in energy, system A can be found in states with different energies at different times. The probabilities of finding system A in the kth state with energy E_k are denoted $P^A = [p_1, p_2, \ldots p_k, \ldots, p_N]$. We define the average energy $U_A = \langle E \rangle_A$ and $U_B = \langle E \rangle_B$ for each system.

The key assumption connecting information theory to physical systems is this: We assume the probability of finding the physical system in the state E_k is the same as the probability of randomly selecting from a set of E_k's with a probability distribution which maximizes the SMI for each system, given the constraints. Here that the constraint is that the average energy is U.

The probabilities for each *physical* system are therefore given by the Boltzmann probability distribution, Eq. (36), which we derived from applying Jaynes Principle, a purely information theoretic result.

$$p_{k_A}^{(A)} = \frac{e^{-\lambda_1^{(A)} E_k^{(A)}}}{\sum_{k_A} e^{-\lambda_1^{(A)} E_{k_A}^{(A)}}} \tag{47}$$

$$p_{k_B}^{(B)} = \frac{e^{-\lambda_1^{(B)} E_k^{(B)}}}{\sum_{k_B} e^{-\lambda_1^{(B)} E_{k_B}^{(B)}}} \tag{48}$$

Note that λ_1 has the units of inverse energy. From (27) we have

$$U_A = \sum_{k_A} p_{k_A} \; E_{k_A}^{(A)} \tag{49}$$

$$U_B = \sum_{k_B} p_{k_B} . \; E_{k_B}^{(B)} \tag{50}$$

We now *define* the thermodynamic entropy $S(U)$ of each physical system, A or B, as $k_B \log(2)$ times the maximal SMI.

$$\boxed{S(U) \equiv k_B \log(2) \; \mathrm{SMI}^{\max} \equiv k_B S_{\mathrm{MI}}^{\max}} \tag{51}$$

The entropy S is a thermodynamic quantity defined at equilibrium. The SMI by contrast can be calculated for any probability distribution whatsoever. In words, (51) says:

> The value of the thermodynamic entropy $S(U)$ is $k_B \log(2)$ times the amount of missing information in the probability distribution that maximizes the (information theoretic) SMI, given the constraint that the average energy is U.

The entropy is a so-called state function. It depends on U, the average energy but is not determined by the history of the system prior to coming to equilibrium (we can extend the dependence to other state variables like N and V).

The conversion factor between the SMI (in units of bits) and the entropy S (in units of energy/temperature) is $k_B \log(2)$. We can think of this as the entropy associated with 1 bit of missing information. The factor $\log(2)$ simply converts the base of the logarithms from the bit-oriented $\log_2$ to the more conventional natural logarithm. The Boltzmann factor k_B reflects the historical and convenient choice of a unit for temperature (which we will introduce below) in Kelvins rather than, say, in Joules.

If the system A and the bath B are not strongly coupled together, we can assume that the entropy S (and SMI) for the composite A+B system is the sum of the entropy for each system.

$$S_{AB} = S_A + S_B \tag{52}$$

This would not be true, for example, if the state of B simply mirrored the state of A. This lack of correlation (in Shannon's terms, *mutual information*) is part of what we mean by being a thermal bath or reservoir—it has a vast number of degrees of freedom that are independent of the system degrees of freedom.

Conservation of energy in the composite A+B system gives us the constraint

$$U_{AB} = U_A + U_B. \tag{53}$$

Relying on identification of the thermodynamic entropy with the maximum value of SMI from Jaynes principle, we can apply Eq. (40) to each system:

$$S_A = k_B \log(Z_A) + k_B \lambda_A U_A \tag{54}$$

$$S_B = k_B \log(Z_B) + k_B \lambda_B U_B \tag{55}$$

Consider now a small energy fluctuation ΔU that increases the average energy of A, and therefore must decrease the average energy of B by the same amount.

$$S_{AB} = S_A(U_A + \Delta U) + S_B(U_B - \Delta U). \tag{56}$$

We require that S_{AB} be maximal under this variation and so expand each term to first order.

$$S_{AB} = S_A(U_A) + \left(\frac{\partial S_A}{\partial U_A}\right) \Delta U + S_B(U_B) + \left(\frac{\partial S_B}{\partial U_B}\right)(-\Delta U) \tag{57}$$

Requiring that the first order change be zero then yields the stationary condition:

$$\left(\frac{\partial S_A}{\partial U_A}\right) = \left(\frac{\partial S_B}{\partial U_B}\right). \tag{58}$$

Using Eqs. (54) and (55) to evaluate the partial derivatives, we find

$$\lambda_1^{(A)} = \lambda_1^{(B)}. \tag{59}$$

At this point we *define* the temperature to be inversely proportional to the Lagrange multiplier λ_1 associated with the average energy constraint.

$$\boxed{\frac{1}{k_B T} \equiv \lambda_1} \tag{60}$$

So (59) gives us that in equilibrium between the two systems

$$T_A = T_B. \tag{61}$$

and, using (54) and (55) again, gives us

$$\frac{1}{T} = \left(\frac{\partial S}{\partial U}\right). \tag{62}$$

The Boltzmann distribution and the thermodynamic partition function for each system are then given by

$$\boxed{p_k = \frac{1}{Z} e^{-E_k/k_B T}} \tag{63}$$

where

$$\boxed{Z = \sum_k e^{-E_k/k_B T}.} \tag{64}$$

The average energy U is given by

$$U = \langle E \rangle = -\frac{\partial}{\partial \beta_1} \log(Z) \tag{65}$$

or

$$U = k_B T^2 \frac{\partial}{\partial T} \log(Z) \tag{66}$$

The information theoretic expression in (41) now becomes the definition of the Helmholtz free energy,

$$\boxed{F \equiv -k_B T \log(Z)} \tag{67}$$

and (40) becomes

$$\boxed{F = U - TS.} \tag{68}$$

If we consider differential changes at constant temperature (and volume) we have

$$\boxed{dF = dU - TdS} \tag{69}$$

which is a key thermodynamic identity. At equilibrium we have from (62),

$$\boxed{dU = TdS} \tag{70}$$

or,

$$dS = \frac{dU}{T}, \tag{71}$$

so at equilibrium $dF = 0$, that is, the free energy is at a minimum.

For a system in thermal equilibrium with a large heat bath the free energy is a minimum. A large mechanical system that can dissipate energy minimizes its total energy U, settling down to an energy minimum. The free energy $F = U - TS$ is the corresponding quantity for a system in thermal equilibrium with a heat bath. It reflects the interplay between lowering the energy and thermal excitation.

We note that Eq. (71) is the original Clausius definition of entropy. Thus, starting with the Shannon information theoretic definition of entropy (3) we have arrived at both the Boltzmann expression (25) and the thermodynamic expression of Clausius. The fact that the historical order was exactly opposite to this logical order has been one source of confusion.

External Work The energy of each state of the system may depend on externally imposed parameters like the volume V, applied magnetic field B, an applied electric field, etc. For example, for an ideal gas we take $E_k = E_k(V)$, then seek the average value of the differential energy shift with volume. The shift is fundamental because the single-particle quantum energy levels move up in energy as the volume is decreased. Classically, it suffices for us to note that a piston must apply a force in the direction of its motion to squeeze a gas into a smaller volume, thus doing positive work on the gas. First, we take the derivative of $\log(Z)$ with respect to the volume.

$$\frac{\partial}{\partial V}\log(Z) = \frac{\frac{\partial}{\partial V}\sum_k e^{-E_k/k_BT}}{\sum_k e^{-E_k/k_BT}} = -\frac{1}{k_BT}\left\langle \frac{\partial E_k}{\partial V}\right\rangle \tag{72}$$

This average value of the energy shift, using the probabilities in (63) is then defined to be the pressure.

$$p \equiv \left\langle -\frac{\partial E_k}{\partial V}\right\rangle \tag{73}$$

The minus sign is because when the volume is decreased the energy increases as does the pressure. Using (72) we then have:

$$\boxed{p = -\left(\frac{\partial F}{\partial V}\right)_T.} \tag{74}$$

Similarly, for the expectation value of the magnetization

$$M = -\left(\frac{\partial F}{\partial B}\right)_T \tag{75}$$

and so forth.

Applying an external force to the system mechanically, electrically, or magnetically is another way to increase the average energy of the system. Equation (73) can be written in terms of this shift.

$$dU = -pdV \tag{76}$$

This kind of direct transfer of energy from outside the system to, or from, the system is called *work*. So there are two ways the average energy of the system can be changed: by heat transfer or by work. Heat is the transfer of a certain amount of energy Q accompanied by a change in entropy. To include both kinds of energy change, we need to modify (70) to:

$$dU = dW + dQ. \tag{77}$$

This is the first law of thermodynamics, the conservation of energy.

In the case of compressing the volume with a frictionless piston, for example, $dW = -pdV$. We will be concerned in Sect. 5 with doing electrical work. If a voltage source transfers differential charge dq across a voltage difference V then, neglecting the resistance of conductors, it does work

$$dW = Vdq. \tag{78}$$

We will interpret the symbol V as voltage or volume by context.

4.2 Statistical Mechanics of the Grand Canonical Ensemble

We can extend the application of the information theoretic results of the previous section to the grand canonical ensemble by considering a system and bath both comprised of particles. Up to now we did not need that assumption so the results have even broader applicability. If, in addition to energy, particles can flow between the system and the bath, we can label states of the system with both energy E_k and the number of particle N_k. The number of particles can fluctuate and in equilibrium we have a constant expectation value $\langle N \rangle$. This constraint gives us a distribution of the form in Eq. (44), where we now identify $\lambda_2 = -\mu/k_B T$, defining the chemical potential μ,

$$p_k = \frac{e^{-(E_k - \mu N_k)/k_B T}}{\sum_k e^{-(E_k - \mu N_k)/k_B T}} = \frac{e^{-(E_k - \mu N_k)/k_B T}}{Z_G} \tag{79}$$

The constraint that the average particle number is $\langle N \rangle$ gives us, in analogy to (38):

$$\langle N \rangle = -\frac{\partial}{\partial \lambda_2} \log(Z_G) \tag{80}$$

The corresponding free energy expression is obtained by the same procedure that connected (40) and (41) to the free energy (67), now using (46) to yield

$$F = -k_B T \log(Z_G) \tag{81}$$

and

$$\boxed{F = U + \mu \langle N \rangle - TS.} \tag{82}$$

The chemical potential μ is the driver for particle exchange between systems in diffusive contact.

Non-interacting Fermions and Bosons In the special case of a group of non-interacting particles in thermal and diffusive contact with a reservoir at temperature T, we can obtain the standard results for fermion and boson statistics using the probability distribution (79), derived from the information theoretic result (44), and the basic rules for state occupancy. For this case we consider a fixed set of single-particle energy levels e_i with the ith level occupied by n_i particles. The total energy of a particular configuration of occupancies will be

$$E = \sum_i n_i e_i \qquad N = \sum_i n_i \tag{83}$$

For fermions, we need only the fact that each level can have occupancy of either 0 or 1 but no greater. The partition function can be written in a factored form with each factor corresponding to the possible occupancies of each level.

$$Z_G = \left(1 + e^{-(e_1 - \mu)/k_B T}\right)\left(1 + e^{-(e_2 - \mu)/k_B T}\right)\left(1 + e^{-(e_3 - \mu)/k_B T}\right)\cdots \tag{84}$$

$$= \prod_i \left(1 + e^{-(e_i - \mu)/k_B T}\right) \tag{85}$$

The average occupancy of the jth level is then given by:

$$\langle n_j \rangle = \frac{\left(0 + 1e^{-(e_j-\mu)/k_BT}\right) \prod_{i \neq j} \left(1 + e^{-(e_i-\mu)/k_BT}\right)}{\prod_i \left(1 + e^{-(e_i-\mu)/k_BT}\right)} \tag{86}$$

$$= \frac{e^{-(e_j-\mu)/k_BT}}{1 + e^{-(e_j-\mu)/k_BT}} \tag{87}$$

$$\boxed{\langle n_j \rangle = \frac{1}{e^{(e_j-\mu)/k_BT} + 1}}. \tag{88}$$

Equation (88) is, of course, the famous Fermi-Dirac distribution function.

The factorization of the partition function and subsequent cancellation is a general feature of composites of non-interacting systems. Each single-particle energy level acts like a separate system in thermal and diffusive equilibrium with the reservoir.

For bosons, each level can be occupied by any number of particles.

$$\langle n_j \rangle = \frac{e^{-(e_j-\mu)/k_BT} + 2e^{-2(e_j-\mu)/k_BT} + 3e^{-3(e_j-\mu)/k_BT} + \cdots}{1 + e^{-(e_j-\mu)/k_BT} + e^{-2(e_j-\mu)/k_BT} + e^{-3(e_j-\mu)/k_BT} + \cdots} \tag{89}$$

The denominator is a geometric series in $x = e^{-(e_j-\mu)/k_BT)}$ yielding $1/(1-x)$. In terms of the same x, the numerator is $S = x + 2x^2 + 3x^3 + \cdots$. We note $S - xS = x + x^2 + x^3 + \cdots$, which is the geometric series minus 1 or $S = x/(1-x)^2$. We arrive at the Bose-Einstein distribution function:

$$\boxed{\langle n_j \rangle = \frac{1}{e^{(e_j-\mu)/k_BT} - 1}}. \tag{90}$$

4.3 Exploring the System Microstates

Consider a monatomic classical ideal gas (no interactions between particles) with N particles of mass m in a volume V at temperature T in the dilute limit. The dilute limit is when the density of particles is low enough that the average occupancy of each energy level is much less than 1. In this case it makes no difference whether the particles are fermions or bosons. A noble gas is well approximated this way. The thermodynamic entropy S and the corresponding SMI are given by the Sakur-Tetrode equation:

$$\text{SMI} = \frac{1}{k_B \log(2)} S(N, V, T) = N \log_2 \left[\frac{V}{N} \left(\frac{m k_B T}{2\pi \hbar^2} \right)^{3/2} \right] + \frac{5}{2} N \tag{91}$$

This equation can be derived from the Jaynes Maximum Entropy principle [5] or from standard thermodynamics. Though describing a classical gas, the expression contains Planck's constant because it's necessary to enumerate the smallest volumes in phase space (limited by the uncertainty relationship) to give an absolute number.

For a liter of Argon gas at standard temperature and pressure, the SMI from (91) is about 10^{23} bits. Recall that this is the average length of the binary register necessary to specify all the accessible microstates of the gas.[7] The number of possible microstates is therefore

$$N_{\text{microstates}} \approx 2^{\text{SMI}} = 2^{(10^{24})} \approx 10^{(10^{23})}. \tag{92}$$

Let us imagine this liter of argon moving from one accessible microstate to another according to its internal dynamics. We can imagine a binary register holding the (compressed) index of the current microstate of the gas which keeps clicking from one number to the next as the gas particle move from state to state. How much time does it take to go from one state to another? Well, to change states requires the atomic nuclei to move positions. Suppose the shortest time to move nuclear positions establishes the "tic" of the clock at which point the microstate register changes to record the next microstate index. We are looking for an *upper bound* to the number of microstates explored in a given time, so we will take the shortest possible clock tic, the light travel time across a nucleus $T_{\text{tic}} \approx 10^{-25}$ s. The nuclei could hardly have changed positions faster than that. How many microstates could the liter of argon have explored? The time since the big bang is $T_{\text{universe}} \approx 10^{18}$ s. An upper bound on the number of microstates that the liter of argon could possibly have explored since the beginning of the universe is then

$$N_{\text{microstates explored}} \leq \frac{T_{\text{universe}}}{T_{\text{tic}}} = \frac{10^{18}\,\text{s}}{10^{-25}\,\text{s}} = 10^{43}. \tag{93}$$

Therefore, the *fraction* of the accessible microstates that could possibly have been explored since the universe began is

$$\frac{N_{\text{microstates explored}}}{N_{\text{microstates}}} = \frac{10^{43}}{10^{(10^{23})}} = 10^{(-10^{23}+43)} \approx 10^{-(10^{23})}. \tag{94}$$

[7] Note that we assume that state indices (a series of 1's and 0's specifying each particular state) are chosen in an optimal way, employing a so-called Huffman code, that uses fewer bits to specify more probable states and longer bit sequences for rarer states. The average register length is the average index length weighted by the state probabilities.

This is an extremely small number. We conclude that any actual gas is visiting only an extraordinarily small fraction of it possible microstates. The vast majority of terms in the sums like the partition function represent states that *have never* been realized by the system and *will never* be realized by the system.

The connection established by Jaynes using Shannon entropy (SMI) put statistical mechanics on a much firmer footing for both classical and quantum cases. But it fair to wonder: why does this work so well in so many situations? We can view the system averages, like the average energy, as time averages. Or we can view these averages as ensemble averages over many systems with identical macroscopic properties but different microscopic configurations. In either case, the microscopic state of a particular system is the result of its particular *dynamics*, the equations of motion and its past history or initial state. The connection with information theory is fundamentally made by mapping one problem: (a) the physical process of dynamical motion, onto a second problem (b) the statistics of selecting a ball with an energy label on it from an urn with an optimal (maximum SMI) distribution of balls given the average energy.

The number of balls in the urn that corresponds to the possibilities for the physical system is staggeringly large. Whatever sample we use for the average is, in an important sense, a very small sample. We have just seen that the system state space is vastly larger than it could explore in any reasonable time.

The probability distribution is determined by the physical dynamics (classical or quantum) and yet those dynamics do not matter to statistical mechanics. Statistical mechanics is remarkably independent of mechanics. We know some dynamical model systems pull systems into attractors rather than distributing them uniformly across state space. So the question becomes: what properties of real physical dynamical systems make them so amenable to very accurate description by selecting "typical" states from the incredibly vast state space available? This is an active research area. An example of sustained and sophisticated work on this issue in the quantum thermodynamics is the work of Gemmer, Michel, and Mahler [8].

5 The Landauer Principle

The Landauer Principle (LP) asserts that for a physical system representing an information state, loss of one bit of information necessarily entails dissipation to the environment of a minimum amount of heat equal to $k_B T \log(2)$. If information is not lost, there is no minimum amount of heat dissipation necessary.

Any logically irreversible operation, AND, OR, SUM, etc., involves a loss of information in the sense that inputs cannot be logically inferred from the outputs. The archetypal irreversible operation is erasure, so we will focus our attention on that.

For specific devices the heat dissipation may, of course, be much more than the fundamental minimum. Modern CMOS transistors operate with orders of magnitude more energy dissipated by each transition. If the device and associated architecture

is designed optimally (adiabatic logic), it may be possible to lower the dissipation by switching its state more smoothly and slowly. The Landauer Principle places a fundamental lower limit on how much heat dissipation must occur, depending on the amount of information that is lost.

We discuss three arguments for the Landauer Principle. The first, the many-to-one argument is the one Landauer himself usually employed, though he most often simply asserted the principle as self-evident. The second grounds the argument on the Second Law of Thermodynamics. The third is a calculation on a minimal model system. This has the advantage of being a mathematical result with clear assumptions and clear conclusions. Of course it is susceptible to the objection that there might be another way of constructing a memory that would violate LP. Be that as it may, it is very clarifying to see LP simply emerge from a simple calculation. It also forces the issue of how to define the entropy of memory states that cannot be equilibrium states, using the same approach as in Sect. 4.

5.1 The Many-to-One Argument

This argument for LP is based on the time-reversal symmetry of the physical law at the microscale. We will consider a physical system that has three states (or regions of state space) that encode a binary 1, binary 0, or a null state holding no information. Why not just use the two binary states? Choosing to define erasure as "set-to-zero" results in a morass of ambiguity because we cannot physically distinguish the erasure process from the physical process of writing a 0. Consider the erasure of a single bit of information that is represented by the physical system. Let us assume that we do not know the value of the initial bit, 1 or 0, but need to create with *one protocol*, a series of externally controlled operations, that will result in the physical system being set to the null state.

This is shown schematically in Fig. 6. The proposed protocol would be such that given an initial physical configuration of the system that corresponds to *either* a 0 or a 1 bit, the protocol would be such that the physical system would evolve in time under the relevant physical law to the state representing the null bit, as shown in the figure.

But, the LP arguments objects, something like Fig. 6 cannot occur. If it did, then one could start at the null state, and run the protocol backwards. But which state would then result, 1 or 0? Running a movie of the whole physical process backward should be a valid physical process.[8] Fundamental physics takes one (classical or quantum) state forward in time to a unique future quantum state.

[8]The weak interaction responsible for the decay of the neutral B meson has been directly shown to violate time reversal symmetry. See J. P. Lees et al., "Observation of Time-Reversal Violation in the B^0 Meson System," Phys. Rev. Lett. 109, 211801 (2012). We will restrict our considerations to systems not involving B or K mesons, or the weak interaction.

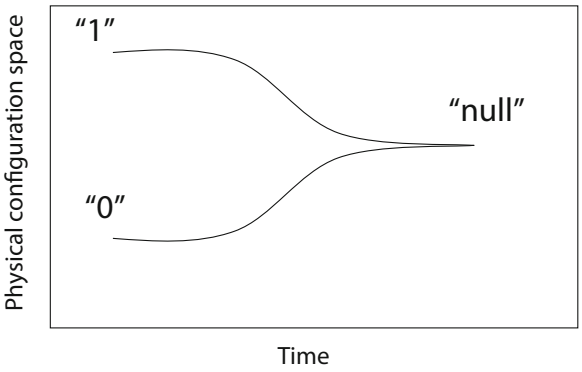

Fig. 6 The many-to-one quality of bit erasure to a null state. A general-purpose physical erasure procedure would have to be able to take either a 1 or a 0 state and move it into the null state. It would have to be the same procedure irrespective of the initial state and so work on an unknown bit. Reversing the temporal sequence of the procedure should be possible because the microscopic laws of physics are reversible and one-to-one. There cannot therefore be a unique backward evolution that would restore the state to its original 1 or 0

Therefore, if it *looks* like the situation of Fig. 6 is occurring, careful examination will reveal that there is at least one other system involved that is not being properly accounted for. As shown in Fig. 7a, there must be at least one other degree of freedom in the process, shown in the figure as the z-coordinate that, is different between the two outcomes. The figure shows the z-coordinate of this auxiliary system initially at 0, and then moving to ± 1 depending on the initial state of the system. The auxiliary degree of freedom could be, for example, another system which is initially in the null state and then put by the protocol in the same 1 or 0 state as the primary system was initially. Of course, it could store the inverse state as well. In this case the information of the enlarged system, including the original system and the copy, has not lost the original information.

The auxiliary system could also be a very large complex system like a heat bath. In that case the bath, taken as a whole, retains, in an entirely inaccessible way, a copy of the original bit. Again, the enlarged system+bath has not lost the information fundamentally, thus preserving the time-reversibility of the system. There are two different final states for the bath, one in which it has interacted with the system in a 1 state and another in which it has interacted with the system in the 0 state. The SMI of the bath has now increased because, knowing only the macrostate $(P, V, T, \ldots)$ of the bath, the information about the which bit was originally stored in the system is missing. If the bath is at thermal equilibrium at temperature T, the increase in thermodynamic entropy must be $\Delta S = k_B \log(2)$ with the corresponding heat transfer to the environment $\Delta Q = T\Delta S = k_B T \log(2)$.

If we deliberately make a copy of the bit, physics does not prevent us from exploiting the fact that we have a copy to create different erase-one and erase-zero

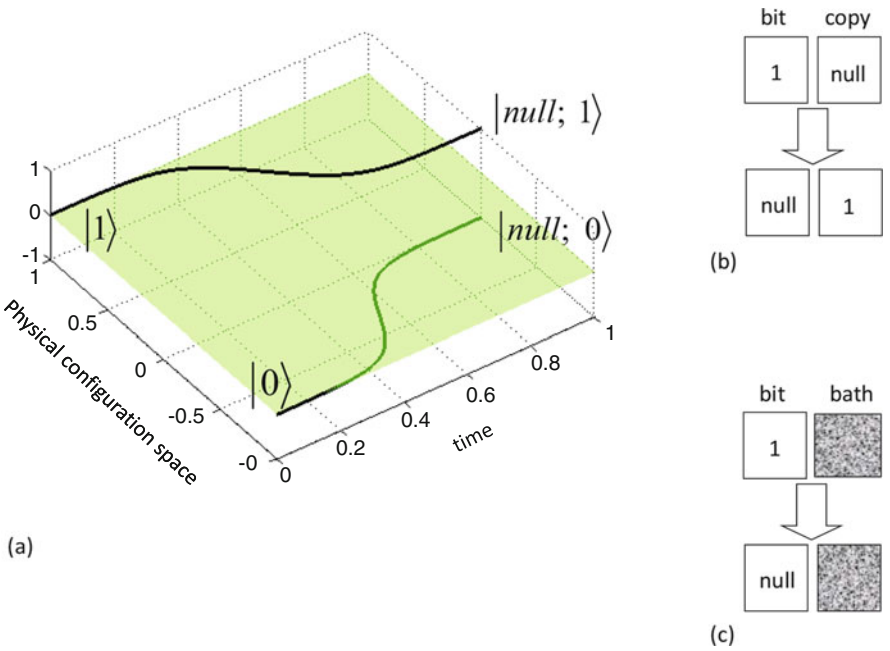

Fig. 7 The reality of an apparent many-to-one bit erasure process. (**a**) If a system *appears* to evolve as in Fig. 6 under the reversible laws of physics, there must actually be another physical system coupled to the bit-storing system. The physical state of the copy system is here represented on the z-axis of the graph. (**b**) The auxiliary bit could be simply another physical system that makes a copy of the original bit, for example a neighboring bit or one in a measurement apparatus. Having a copy allows there to be a different physical process to erase a 1 than to erase a 0, shown as separate curves in the figure, so the process is reversible. (**c**) Alternatively, the copy could be contained in the large system of a thermal bath in contact with the bit system. The copy in that case is encoded in the many degrees of freedom of the bath and is unrecoverable, leading to an irreversible erasure. This process transfers entropy and heat to the bath. After erasure, the increased energy of the bath means there are twice as many accessible bath states—those corresponding to the 1 bit having been erased, and those corresponding to the 0 bit having been erased

protocols. We only are forced to pay the Landauer price if there is no copy, or if there is one and we just fail to make use of it.

5.2 Argument from the Second Law of Thermodynamics

Consider again the small system A in thermal equilibrium with a very large system (a heat bath) B as shown in Fig. 5. We will assume that B has temperature T and that system B is in thermal equilibrium with A and so has the same temperature. We now suppose that there are a set of macroscopic controls that allow us to externally manipulate the state of B. In examples that we will flesh out below,

these will be electrodes whose electrostatic potential we can change. For the often-considered example of a perfect single-particle gas, manipulation is typically by moving pistons and inserting and removing barriers. However it is accomplished, suppose the entropy S_A of A is *reduced* by the equivalent of one bit of information (see Eq. (51)). This is what we mean by erasure.

$$\mathrm{SMI}_A(\text{final}) - \mathrm{SMI}_A(\text{initial}) = -1 \text{ bit} \tag{95}$$

$$S_A(\text{final}) - S_A(\text{final}) = -k_B \log(2) \quad = -\Delta S_{\text{bit}} \tag{96}$$

The heat bath is large so this has no effect on its temperature. We will again assume that the entropy of the global system comprising A and B is the sum of the entropies of the components. The internal motion of the heat bath B is not correlated to the motion of A.

$$S_{AB} = S_A + S_B \tag{97}$$

The manipulations of A we will assume do not perform a measurement of the microstate of B (or of A). We assume that the manipulation does not give us any new information about AB. Therefore, the change in entropy of the global system must be either zero or positive. The amount of information missing about the microstate of the composite system AB can only increase since we (or anyone or anything doing the manipulation) have not reduced the missing information about the microstate of AB. This is the heart of the Second Law of thermodynamics.

$$\Delta S_{AB} = \Delta S_A + \Delta S_B \geq 0 \tag{98}$$

Therefore:

$$\Delta S_B \geq -\Delta S_A \tag{99}$$

$$\Delta S_B \geq k_B \log(2) = \Delta S_{(\text{bit})} \tag{100}$$

After the manipulation is complete, the bath system B is in an equilibrium state with an increased entropy and an energy which is larger by at least the energy corresponding to one bit.

$$dU_B = T dS_B \geq k_B T \log(2). \tag{101}$$

This is, as we have stressed, a fundamental minimum, not a characteristic energy.

5.3 Direct Calculation of Erasure in Minimal System

Representing *encoded information* with the *raw information* of a physical system involves both the dynamics of the physical system and the encoding scheme. The encoding scheme maps areas of the accessible region of the physical state space of the system to specific information states. Figure 8 represents a physical system with three states for a single particle which can be in one of three state (dots) indexed 1, 2, and 3. The figure illustrates (a) the particle on the right representing a "1", (b) the particle on the left representing a "0", and (c) the particle in the middle representing a "null" state containing no binary information.

The encoding scheme for the three logical states can be defined in terms of the probability of the occupancy of each state. The probability of the system being found in state 1, 2, or 3, we denote $[P_1, P_2, P_3]$, and the energy of each state is denoted $[E_1, E_2, E_3]$. We can choose a threshold value P_{th} and encode a binary 1 by a state with $P_3 > P_{\text{th}}$, a binary 0 with $P_1 > P_{\text{th}}$, and the null state with $P_2 > P_{\text{th}}$. If no state has probability above the threshold, the result is not yet a valid state. This is normal in the switching regime. A robust system is designed so the logical state is valid and unambiguous when it is read or copied.

The dots can represent abstract states of the system or literal dots. In the quantum-dot cellular automata (QCA) scheme, they correspond to localized electron states on literal quantum dots. We will treat the system completely classically in this section, and will for convenience assume a single positive charge; the quantum treatment is taken up in Sect. 6. Actual QCA three-dot cells have been fabricated in metal-dot systems and synthesized in single-molecules [9, 10]. The threshold probability P_{th} in these systems is not set simply arbitrarily, but only needs to be large enough for the next stage in logical operations to reset the bit strongly to a logical 1 or 0. Power gain from stage to stage means that P_{th} could be 0.8, for example, and still be strong enough to be effective as transmittable bit information.

We are interested in the process of bit erasure in this system when it is in thermal contact with a heat bath of temperature T. We can control the energy of each state with a set of control voltages capacitively coupled to each dot. The energetic landscape is shown in (a–c) of (8). We choose a $20k_BT$ energy separation between low, active (0 or 1), and high energy state. When state 2 is high, it acts as a barrier to hold the 1 or 0 bit. When state 2 is low it acts as a well to hold the particle in a neutral position. Thus the energy E_2 acts as a clock which can latch a bit by being raised, or erase a bit by being lowered, returning it to the null state.

In the following, we will for convenience simply refer to the fully localized states: $P = [0, 0, 1]$ for the 1 state, $P = [1, 0, 0]$ for the 0 state, and $P = [0, 1, 0]$ for the null state. Examining Fig. 8 reveals an important point:

> A physical memory containing information cannot be in a thermal equilibrium state.

The null state is an equilibrium state satisfying the Boltzmann distribution (63), but neither the 1 nor the 0 state shown in Fig. 8a,b can be. The reason is clear enough—to be a memory is to hold information stored at a previous time. The

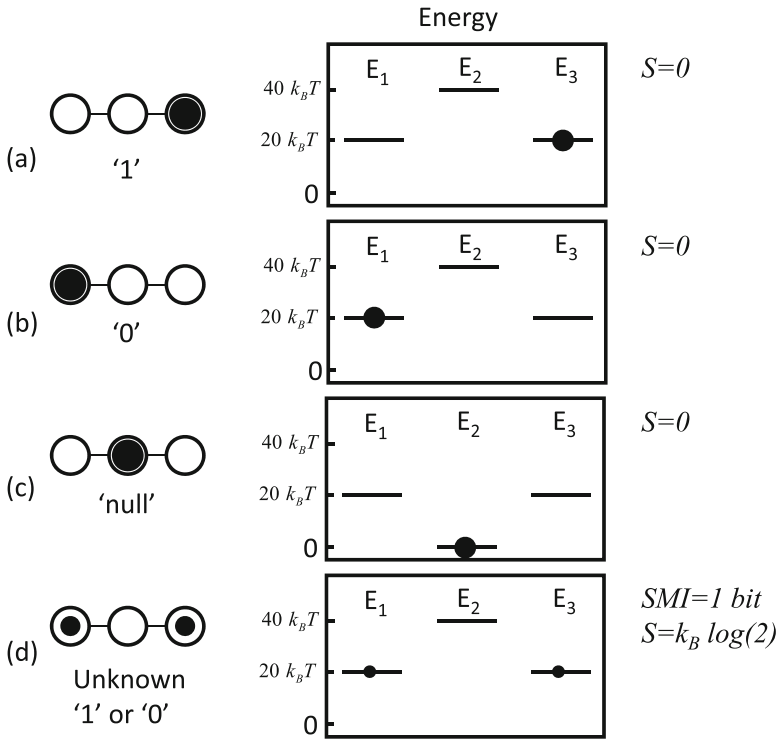

Fig. 8 Encoding information with three physical states. A particle can be in one of three dots encoding (**a**) a logical 1, (**b**) a logical 0, or (**c**) a logical null state. The energy landscape and probability density for each configuration are shown on the right. The solid circles represent the probability of finding the particle on a particular dot. Dot 2 acts as the barrier for holding the particle in the 1 or 0 state in (**a**) and (**b**). It acts as the low energy null state in (**c**). If the bit value is an unknown 1 or 0, as in (**d**), then the energy landscape is the same as in (**a**) or (**b**), but we must assign a probability distribution that is evenly divided between the 1 and 0 states

state of a physical memory *must* depend on the past and not just the current temperature and applied voltages, and whatever macro constraints are relevant. A thermal equilibrium state, by contrast, *cannot* depend on the past, but only on the present conditions.

A physical memory must be in a long-lived metastable state. When E_2 is high it must create a barrier that is sufficiently opaque to hold the particle in the 1 or 3 states for the relevant timescale—microseconds to years. Beyond that, the details of the physical dynamics that allow state transitions between states 1, 2, and 3 do not concern us. The 1 and 0 states of Fig. 8a,b are certainly low energy states—the problem is they preferentially occupy one active dot and not the other though it has the same energy. We will assume in our example that the $20k_BT$ barrier of E_2 for these states is indeed adequate to hold a 1 or 0 bit long enough to be a memory. If a higher barrier was needed, it could be created by raising the potential on dot 2 further.

Entropy for Representing a Known or Unknown Bit For the 1 or 0 states shown in Fig. 8a,b, we must take care to return to our derivation of the thermodynamic entropy S in Sect. 4.1 and now include the fact that we know the value of the bit. This should be treated as an additional constraint in the Jaynes maximum entropy principle as developed in Sect. 3.7.

Consider the case when we know the bit stored is a binary 0. We again require that the thermodynamic entropy be $k_B \log(2)$ times the SMI of that probability distribution which maximizes the Shannon entropy, subject to the given constraints. The constraints are now (a) the probabilities sum to 1, (b) the average energy $\langle E \rangle = U$, and (c) $P_3 = 0$. To find the maximum entropy distribution, we construct the Lagrangian, which now has the additional Lagrange multiplier λ_3.

$$\mathscr{L} = -\sum_k P_k \log(P_k) - (\lambda_0 - 1)\left(\sum_k P_k - 1\right) - \lambda_1 \left(\sum_k P_k E_k - \langle E \rangle\right) - \lambda_3 P_3 \tag{102}$$

The Lagrange equation obtained from requiring the extremum with respect to λ_3, $\partial \mathscr{L}/\partial \lambda_3 = 0$, yields simply $P_3 = 0$ with λ_3 arbitrary. We find the extremum of (102) with respect to all the other P_k's as before. The derivations of all thermodynamic quantities derived in Sect. 4.1 go through as before, simply omitting occupancy of dot 3 as a possibility. The Boltzmann distribution then applies as before to all probabilities but P_3. The thermodynamic entropy is then: $S = k_B \log(2)\mathrm{SMI}([P_1, P_2])$. If the bit was a 1, the constraint would be $P_1 = 0$. In this straightforward way we can apply the definition of thermodynamic entropy to include a state storing a known bit, even though it represents a nonequilibrium metastable state—as it must.

By contrast, the state shown in Fig. 8d represents a reliably stored but *unknown* bit. Since we do not know the value we must assign probabilities $P = [0.5, 0, 0.5]$, the probabilities we would get for the equilibrium state with the barrier high (hence $P_2 = 0$). The associated $SMI = 1$ bit (one bit of missing information) and the thermodynamic entropy $S = k_B \log(2)$. In terms of thermodynamic quantities, it is indistinguishable from an equilibrium state. But because the barrier is sufficient to hold the unknown, but definite, bit for the relevant timescale, it should not be imagined to be switching back and forth. It is not a "bit gas," as Norton has characterized it in [2], but is simply unknown.

The null state is an equilibrium state with entropy $S = k_B \log(2)\mathrm{SMI}[P] = 0$.

Thermodynamic Quantities During Bit Operations We examine below three basic operations: writing a bit, erasing an unknown bit, and erasing a known bit. In each case we will manipulate the three dot energies in time, $E_1(t)$, $E_2(t)$, $E_3(t)$, according to a protocol designed to accomplish the task. Unless otherwise noted, at each point in time, we assume the system is in thermal equilibrium with a bath at temperature T. We assume that the variation in time is at gradual enough that the system is always in its equilibrium state, except as noted for a stored known bit. Therefore, temporal dynamics play no essential role here and we use arbitrary units

spanning the event with $t = [0, 1]$. During the operation we calculate the following four thermodynamic quantities and plot them as functions of time:

1. The equilibrium probabilities $P = [P_1(t), P_2(t), P_3(t)]$ for finding the particle on each dot.

$$P_i(t) = \frac{e^{-E_i(t)/k_B T}}{\sum_k e^{-E_k(t)/k_B T}} \tag{103}$$

   The corresponding expectation value of the charge on the dot is $q_i(t) = P_i(t)e$, where e is the elementary charge. As discussed above, for a state representing a known bit, we simply set P_1 or P_3 to zero.
2. The thermodynamic entropy is given by:

$$S(t) = -k_B \log(2) \sum_k P_k(t) \log(P_k). \tag{104}$$

3. The cumulative amount of heat transferred ΔQ_{bath} to the thermal bath from the beginning of the operation until the time t.

$$\Delta Q(t) = -\int_0^t dQ = -\int_0^t T dS = T(S(0) - S(t)) \tag{105}$$

   The sign is chosen so that net heat flowing from the system to the bath is positive, and net heat flowing from the bath into the system is negative.
4. The work done on the system by the external control electrodes. The electrical potential of each dot i is $V_i(t) = E_i(t)/e$, and the differential work done by the external circuit is

$$dW_i(it) = -V_i(t) dq_i(t). \tag{106}$$

   The minus sign is because when the external circuit raises the dot energy, it actually decreases the dot charge because the thermal occupancy of the dot is reduced via (103).

$$W(t) = \sum_i \int_0^t dW_i(t') = -\sum_i \int_0^t V_i(t') \frac{dq_i(t')}{dt'} dt' \tag{107}$$

   This does not include the work done by the voltage sources on the gate electrodes that are capacitively coupled to the dots. That is most of the work that the external circuit does, pushing charge and off the gate capacitors. This motion is dissipationless if we neglect the residual resistance of conductors; gradual charging and discharging a capacitor can be done quasi-adiabatically. In any case, dissipation in the gating circuit is not what we are interested in here. Therefore,

the only contributions to W are when charge flows on or off one of the dots in the system itself.

The bit operation is determined by the temperature T and the control protocol defined by $[E_1(t), E_2(t), E_3(t)]$. The calculations using (103)–(107) give us in $P(t)$, $S(t)$, $\Delta Q(t)$, and $W(t)$.

Writing a bit. Starting from the null state we write a 0 bit using the protocol shown in Fig. 9. The potential energy of dot 3 is raised first. Then the energy of dot 2, which holds the particle initially, is ramped up smoothly. As E_2 crosses E_1, Fig. 9c, the particle transfers to dot 1. When E_2 reaches the high energy state, E_3 can be lowered and the particle is held in the state representing a bit value of 0, as shown in Fig. 9d,e. This is, as discussed above, a long-lived metastable state in which E_2 acts as a barrier trapping the particle in dot 1 forming a one-bit memory.

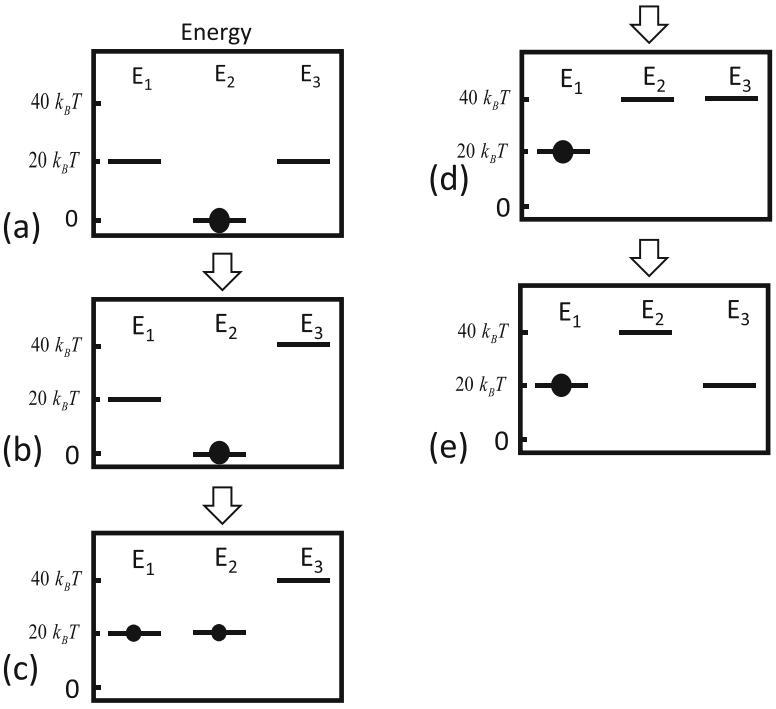

Fig. 9 Protocol for writing a 0 bit. (**a**) The initial state is null, with the particle on dot 2. (**b**) Energy E_3 is raised, biasing the system toward the 0 state. (**c**) The null state energy E_2 is smoothly ramped up, passing the energy E_1. At this point the probability of occupying dots 1 or 2 is equal. (**d**) The energy E_2 is now high and the particle is localized in dot 1, representing a 0 bit. (**e**) E_3 is lowered again and the particle is held on dot 1 by the barrier. This is a memory storing a 0 bit in a long-lived metastable state. A memory-storing state cannot be an equilibrium state; it must depend on the past

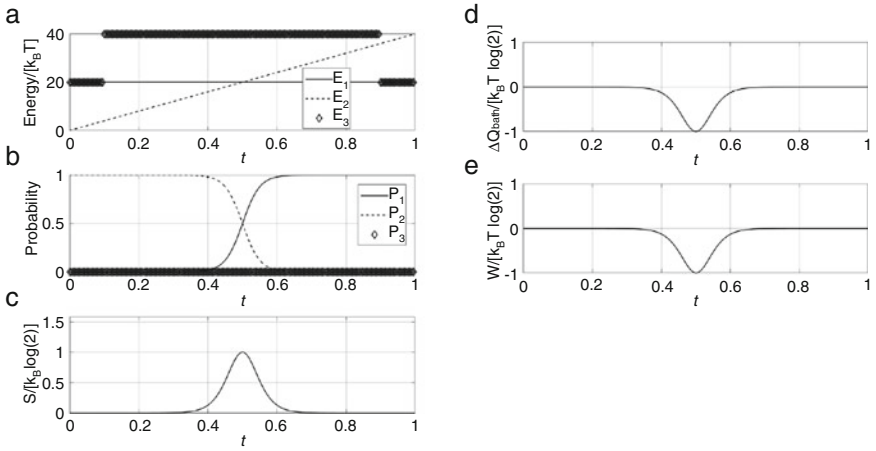

Fig. 10 Thermodynamic quantities during the process of writing a 0 bit using the protocol of Fig. 9. Time is in arbitrary units from 0 to 1. The system is in thermal equilibrium from $t = 0$ to $t = 0.9$, after which it is in a metastable memory state storing a 0 bit. (**a**) The dot energies E_1, E_2, E_3 are shown as functions of time. At $t = 0.1$, E_3 is raised to bias the system toward the 0 state. The energy of dot 2 is ramped up throughout the process, eventually forming a high barrier to hold the stored bit information. At $t = 0.9$ the bias is removed and the bit is firmly latched by $t = 1$. (**b**) The probabilities for dot occupancy P_1, P_2, P_3 are shown as functions of time. (**c**) The thermodynamic entropy of the system calculated from Eq. (104). The peak occurs as E_2 nears and then passes E_1. Thermal excitations between dots 1 and 2 make the location of the particle 1 bit less certain. The peak corresponds to the moment shown in Fig. 9c when $E_2 = E_1$. (**d**) The net heat transferred to the bath ΔQ_{bath} up to time t, calculated from Eq. (105). As E_2 approaches E_1, heat is drawn from the bath; as E_2 moves above E_1, the heat energy is returned to the bath. (**e**) The net work done by the control circuit on the system W, calculated from Eq. (107). When the write process is complete, no net work has been done. This, of course, neglects any heat dissipated within the control circuit that changes the dot energies (e.g., due to nonzero resistance of conductors)

Figure 10a shows the energies $E_1(t)$, $E_2(t)$, and $E_3(t)$. The probabilities of occupancy of each state are shown in Fig. 10b. At each time $0 < t < 0.9$ the probabilities are thermal equilibrium values given by (103). For $0.9 < t < 1$, the system is in the nonequilibrium metastable state where $P_3 = 0$ by assumption. Figure 10c shows the thermodynamic entropy $S(t)$ in units of $k_B \log(2)$ (equivalent to the SMI). As the levels E_1 and E_2 cross, the entropy increases because there is less information about which dot the particle is on. At the crossing point the missing information is 1 bit. What we do not know is the details of the momentary thermal fluctuations which have put the system in state 1 or state 2. Figure 10d shows the heat transferred to the environments $\Delta Q(t)/(k_B T \log(2))$ calculated from (105). As the crossing point is approached the system takes energy from the environment to excite thermal occupancy in the higher energy dot (here dot 1). That energy (heat) is returned to the environment as E_2 continues to increase. Figure 10e shows the work done on the system from

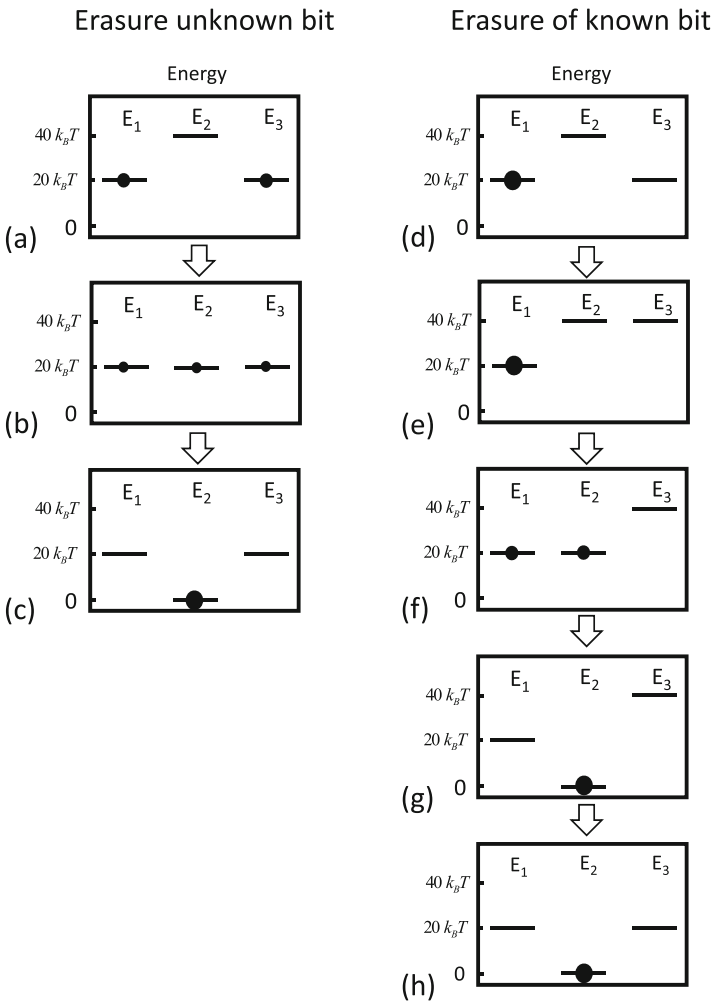

Fig. 11 Erasure protocols for known and unknown bits. Diagrams (**a**)–(**c**) represent the case of an unknown bit and diagrams (**d**)–(**h**) represent the case of a known bit. (**a**) If the bit value stored is unknown, the probability of dot occupancy of dots 1 and 3 is equal. Since we do not know in which direction to bias the system, we can only lower E_2 smoothly in time. (**b**) As E_2 is lowered it passes E_1 and E_3. When they are degenerate the particle could be in any of the three dots. (**c**) Finally the middle state has captured the occupancy and the system is in the null state. (**d**,**e**) For the known bit, here shown to be 0, the system is biased into the state it is already in by raising E_3. (**f**) As E_2 is lowered it passes E_1. (**g**) The system is now in the low energy neutral state. (**h**) The bias can then be removed

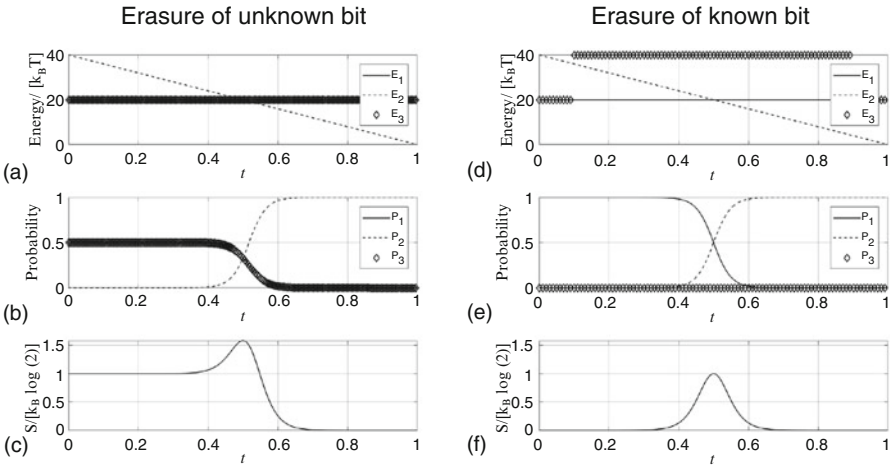

Fig. 12 Energy levels, probabilities, and entropy during bit erasure. Plots (**a**)–(**c**) are for the case of an unknown bit and (**d**)–(**f**) are for a known bit. (**a**) For an unknown bit, E_2 is simply lowered to cause the transition to the null state. (**b**) The probabilities for occupancy of dots 1, 2, and 3 are shown. Initially $P_1 = P_3$ because the bit value is unknown. (**c**) The entropy of the system is initially 1 bit. At the point where all three energy levels are degenerate $S/(k_B \log(2)) = \log_2(3) \approx 1.58$. The entropy then decreases as the system localizes completely on dot 2. The net decrease in entropy is 1 bit. (**d**) For a known bit, the system can be biased into the state it is in, by raising E_3 in this case, until the switching is complete. (**e**) The probabilities for occupancy of dots 1, 2, and 3 are shown for the case of a bit known to be initially 0. Initially only P_1 is nonzero because the system is known to be in the 0 state. (**f**) The entropy for the case of a known bit increases around the level crossing of E_1 and E_2, but is initially and finally 0

0 to t by the external circuit $W(t)/(k_B T \log(2))$ calculated from (107). This is initially negative because energy is being drawn in from the thermal bath, but nets to zero as the energy is returned.

Erasing an unknown bit. Figure 11a shows the energy states for a stored bit. Suppose that this is an unknown bit—there is no other physical copy of the bit that we can use to bias the system into the state it already in. Therefore, all we can do to erase the bit, i.e., set it to the null state, is to lower the barrier E_2 until it localizes the charge in the middle dot. We lower it smoothly, crossing the active-state energies, Fig. 11b, and arrive at the null state shown in Fig. 11c. Again, the system is assumed to be always in thermal equilibrium with the heat bath.

Figure 12a shows the dot energies as a function of time and Fig. 12b shows the probabilities. Initially both P_1 and P_3 are $1/2$ and finally $P_2 = 1$. The entropy shown in Fig. 12c is initially one bit, $k_B \log(2)$. The entropy increases as E_2 passes E_1 and E_3. At the crossing point the SMI is equal to $\log_2(3)$ because each of the three states is equally probable. The entropy drops to zero as the particle becomes completely localized in dot 2. The erasure of an unknown bit involves a lowering of the system entropy by 1 bit. Figure 13a shows the heat transfer to the environment calculated from (105) for this erasure process. It is initially zero

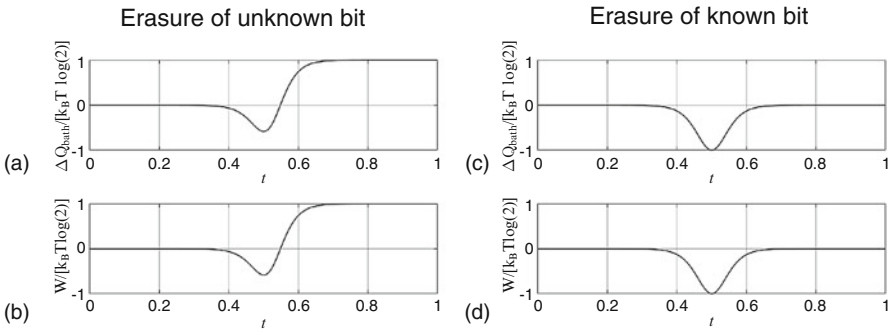

Fig. 13 Heat transfer and work during bit erasure. For the same two protocols of switching as in Figs. 11 and 12, two calculated thermodynamic quantities are shown. Plots (**a**) and (**b**) are for the case of erasure of an unknown bit, and plots (**c**) and (**d**) are for erasure of a known bit. (**a**) For erasure of an unknown bit, the plot shows the net heat transferred to the bath ΔQ_{bath} up to time t, calculated from Eq. (105). As E_2 is lowered, heat is drawn *from* the bath as thermal fluctuations excite the system from E_1 or E_3 to E_2. As E_2 moves below the level of E_1 and E_2, heat energy flows out *to* the bath as the system de-excites and occupies E_2. The net heat transferred to the environment (i.e., dissipated) in this case is $k_B T \log(2)$. (**b**) The net work done by the control circuit on the system W, calculated from Eq. (107), for the case of unknown bit erasure. This is the source of the energy which ends up dissipated as heat to the thermal bath. (**c**) In the case of a known bit, switched according to Fig. 11(**d**)–(**h**), heat drawn in from the bath as E_2 approaches E_1 is returned to the bath as the system moves into the null state. The knowledge of the existing bit state affects the amount of heat dissipated to the bath precisely because it permits a different erasure protocol to be operated. (**d**) Similarly, the net work done by the control circuit on the system W, calculated from Eq. (107), is zero by end of the switching event

and then becomes negative as E_2 is lowered and thermal excitation increases the probability of dot 2 being occupied. The heat transfer swings positive as E_2 drops below the crossing point. The final net value of the heat transferred to the environment is the Landauer Principle limit of $k_B T \log(2)$. Where did this energy come from? It came from the external circuit as shown in Fig. 13b, calculated from (107). The net work done by the circuit during the erasure process is precisely $k_B T \log(2)$.

Erasing a known bit. Now consider the situation of erasing a known bit. In this case we can execute a different set of operations shown in Fig. 11d–h. Figure 11d shows the dot energies when the system holds a known 0 bit (in a metastable rather than equilibrium state). Because it's known to be a 0, we can raise the energy of the other state by increasing the potential energy of dot 3 as shown in Fig. 11d. Then the energy of dot 2 can be lowered as in Fig. 11f,g. Finally, the energy of dot 3 is lowered to restore the whole configuration to the null state.

The energy levels, probabilities, and entropy for this process are shown in Fig. 12d–f. During the initial time $0 < t < 0.1$, corresponding to the configuration of Fig. 12d, the probabilities are not the equilibrium values but rather correspond to the non-equilibrium state of the particle known to be in dot 1. Thereafter thermal equilibrium is assumed at all times. The difference

between the initial probabilities in Fig. 12b and e is precisely the difference between knowing the value of the bit stored and not knowing it. The entropy in known-bit erasure, Fig. 12f, rises as E_2 approaches E_1, but then falls to zero again as the particle is confined on dot 2. The reversible heat transfer shown in Fig. 13a is similar; the system draws in heat from the bath as E_2 comes close enough for thermal excitations from dot 1 to dot 2. But as the system localizes by several $k_B T$, the heat is returned to the bath. The work done on the system, calculated from (107) and shown in Fig. 13a , nets to zero by the end of the switching. This is entirely consistent with the experimental result that erasing a known bit can dissipate orders of magnitude less than $k_B T$ [11].

The issue of a "known" versus "unknown" bit of course has nothing to do with consciousness. "Known" simply means there is another physical representation of the bit that can be used to bias the system toward the state it's already in, as in Fig. 9e. This could be accomplished by a circuit, a computer, or a human brain. In QCA shift registers, the bias is accomplished by having the neighboring bit holding a copy of the target bit [12]. The neighbor is Coulombically coupled to the target bit so it provides an electrostatic bias. Needless to say, one could have a copy of the bit but *not* use it to employ the optimum erasure protocol. In that case one would dissipate at least $k_B T \log(2)$ amount of heat for each erasure.

Critics sometimes ask: How can "subjective" knowledge of the bit state change a physical quantity like heat dissipated? The answer is simply because knowing the existing state allows us to use a different protocol for the erasure, one which initially biases the system into the state it is already in. In the case illustrated in Figs. 9, 10, 11, 12 and 13, we initially raise E_3 to the high level to erase a known 0 bit, but would initially raise E_1 to a high level to erase a known 1 bit.

The free energy $F = U - TS$ is similarly higher for a known state, because S is lower, than for an unknown state with higher S. The Szilard engine is an example of exploiting knowledge of a system state (in that case a single molecule gas) to draw more energy out of the system [13]. A recent physical realization of this sort of "information engine" was reported in [14].[9]

Figure 13a is a clear demonstration of the Landauer Principle in the simplest single-particle system. It requires only Eqs. (103)–(105) above, that is, the Boltzmann thermal occupancy of energy states, the careful definition of entropy for memory states, and the calculation of heat flow. The dissipation of $k_B T \log(2)$ as heat to the bath is unavoidable.

The erasure of an unknown bit above is at each point in time an equilibrium process. It is not, however, time-reversible. Going backwards in the sequence shown in Fig. 11a,b,c would result in latching a *random* bit, determined by a momentary bath fluctuation, not the unknown but definite bit one started with. Imagine the original unknown bit was the result of a long calculation. It is unknown to us until we read it (and thereby make a copy), but its definite value has genuine information

[9]This sort of engine does not, needless to say, violate the second law of thermodynamics or create a perpetual motion machine of the second kind.

about the calculation's result. The time-reversed protocol would not restore the result, but replace it with a random bit. The probability distribution for an unknown bit holding the important result of a long calculation (e.g., target is friend or foe) is identical with the probability distribution of a meaningless random bit latched from a thermal fluctuation.[10]

By contrast the erasure protocol of Fig. 11d–h is time reversible. It is in fact just the time-reversed version of the writing protocol shown in Fig. 9. Both assume we have another copy of the bit to determine bias applied in the writing sequence (do we raise the potential on dot 1 or dot 3?) and the subsequent erasure.

The main result of this calculation is the difference between the heat dissipated to the environment for erasure of an unknown bit, shown in Fig. 13a, and that for erasure of a known bit, shown in Fig. 13c. The Landauer Principle result is clear—heat energy of $k_B T \log(2)$ is dissipated for erasure of an unknown bit, but there is no minimum dissipation required for a known bit. If the bit is known, of course, there is a copy of the bit somewhere and the information is not truly lost. If the copy is eventually erased, heat dissipation must occur.

6 Quantum Mechanics, Entropy, and the Nature of the Physical Law

6.1 Quantum Formalism and Probabilities

In quantum mechanics a physical system is described by a state vector $|\psi\rangle$, a so-called Dirac "ket" in a Hilbert space.[11] The state vector is a *complete description* of the system. It contains all there is to know about the system—all the physical world itself "knows" about the system.

The inner product between two state vectors is denoted $\langle\phi \mid \psi\rangle$ and gives the probability amplitude that a system in state $|\psi\rangle$ will be measured to be in state $|\phi\rangle$. The probability is the absolute square of the probability amplitude.

$$\boxed{p_\phi = |\langle\phi \mid \psi\rangle|^2} \tag{108}$$

Equation (108) is the Born rule.

Dynamical observables, such as position x and momentum p, are represented in quantum mechanics by Hermitian operators that map one ket onto another

[10] "Meaningless" is, of course, a question of context—maybe it was meant as a probe to measure bath fluctuations.

[11] A Hilbert space is a complex linear vector space with an inner product that produces a norm on the space. Using this norm, all Cauchy sequences of vectors in the space converge to a vector in the space.

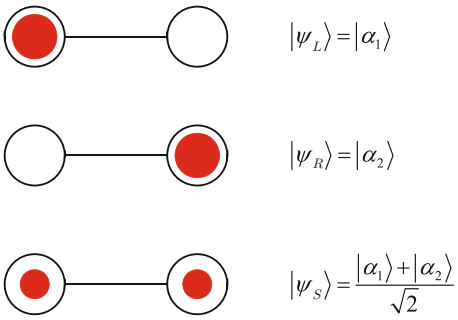

Fig. 14 Quantum state of a particle in two coupled quantum dots. The particle could be in a state that is fully localized on the left dot, $|\psi_L\rangle$, fully localized on the right dot, $|\psi_R\rangle$, or a symmetric superposition of the two $|\psi_S\rangle$. (Any normalized linear combination of $|\psi_L\rangle$ and $|\psi_R\rangle$ is possible.) If the particle is in the superposition state $|\psi_S\rangle$, the result of a measurement of position is not knowable in advance of the measurement. The indeterminacy is not a reflection of the experimenter's ignorance, but is a fundamental feature of the physical world

ket. Measurements of an observable Q always only yield eigenvalues q_k of the associated operator $\hat{Q}$.

$$\hat{Q}\,|\phi_{q_k}\rangle = q_k\,|\phi_{q_k}\rangle \tag{109}$$

The eigenvalue spectrum of an operator may be discrete or continuous. The eigenstates $|\phi_{q_k}\rangle$ are states which have a definite value of the property Q. The probability that a measurement of Q on the system in state $|\psi\rangle$ will yield a specific eigenvalue q_k is therefore:

$$p_{q_k} = \left|\langle\phi_{q_k}\,|\,\psi\rangle\right|^2 \tag{110}$$

This probability is different in kind from the probabilities we have dealt with heretofore. Here the probabilistic nature is not because we lack any information that we could otherwise have. The probabilities here reflect the fact that the physical law and the current state of the system taken together are fundamentally insufficient to determine the result of the measurement. This indeterminacy is a feature of the physical law revealed by quantum mechanics.

6.2 Quantum Mechanical SMI for an Observable

We can define the SMI (Shannon Measure of Information) of the probability distribution for measurements of eigenvalues q_k when the system is in the state ψ.

$$SMI_Q[\psi] \equiv -\sum_k p_k \log_2(p_k) \tag{111}$$

or

$$\boxed{SMI_Q[\psi] = -\sum_k \left|\langle \phi_{q_k} \mid \psi \rangle\right|^2 \log_2 \left(\left|\langle \phi_{q_k} \mid \psi \rangle\right|^2\right)} \tag{112}$$

This represents the amount of missing information (in bits) about the outcome of measurements of Q. This information is missing from the quantum state $|\psi\rangle$ itself. It is not simply missing because of our incomplete knowledge. The physical world itself does not have this information present until measurement is actually made and one particular eigenvalue of $\hat{Q}$ is obtained.

Consider the two-state system illustrated in Fig. 14. A particle can be in the dot on the left, a state represented by $|\psi_L\rangle$, or in the dot on the right, represented by $|\psi_R\rangle$. The symmetric superposition of the two is the state represented by the state

$$|\psi_S\rangle = (|\psi_L\rangle + |\psi_R\rangle)/\sqrt{2}. \tag{113}$$

This superposition state is a unique feature of quantum mechanics. A measurement of the position of the particle in state (113) will always find it in either the left dot or the right dot, each with probability 1/2. We can define an operator corresponding to the observable position of the particle.

$$\hat{X} \equiv |\psi_R\rangle \langle \psi_R| - |\psi_L\rangle \langle \psi_L| \tag{114}$$

The eigenvalues of $\hat{X}$ are $+1$ and -1 corresponding to the particle on the right or on the left. The SMI for this operator corresponds to the amount of missing information about the position of the particle when it's in each of these states.

$$\begin{aligned} SMI_X[\psi_L] &= 0 \\ SMI_X[\psi_R] &= 0 \\ SMI_X[\psi_S] &= 1 \text{ bit} \end{aligned} \tag{115}$$

In the first two cases, the position is definite and there is no information about it that is missing. For the symmetric superposition state, knowing the state tells us nothing about the position. There is 1 bit of position information missing, even though the quantum mechanical description of the state is complete.

Note that the SMI depends on the choice of observable as well as the state itself. For the same state $|\psi_S\rangle$, we could consider the parity operator:

$$\hat{\Pi} \equiv |\psi_L\rangle \langle \psi_R| + |\psi_R\rangle \langle \psi_L| . \tag{116}$$

The state $|\psi_S\rangle$ is an eigenstate of Π with eigenvalue 1—it has a definite value of parity. So there is no missing parity information and SMI_Π is 0. The SMI is basis dependent; it is not the eigenvalue of a Hermitian operator.

We return to considering the information the quantum state gives us about the particles position. It is helpful to ask the question in terms of the amount of information provided by the physical law. Let us suppose that, given an initial state of the system and a precise description of all the potentials, fields, and interactions that are subsequently applied to the particle, the physical law prescribes that at a particular time the system will be in the state $|\psi_L\rangle$. We may then ask: *How much information about the particle's position (by which we mean the results of a measurement of position) does the physical law yield?* The answer is 1 bit. For the state $|\psi_R\rangle$ the physical law also yields 1 bit of information, completely determining where the particle will be found (on the right). But if the physical law tells us that the particle is in the state $|\psi_S\rangle$, it gives us 0 bits of information about the position that will be measured. The information provided by the physical law (evaluating Eq. (9)) is zero. In general, if there are N eigenvalues of the operator $\hat{Q}$, then the physical law gives us a finite amount of information I about the outcome of a measurement of Q, where

$$I[\psi] = \log_2(N) - \text{SMI}_Q[\psi] \quad \text{bits.} \tag{117}$$

Continuous Eigenvalues Consider a wavefunction defined on a *continuous* range of positions. Let $\psi(x)$ be the probability amplitude for finding a particle at position $x \in [0, L]$. If the state for which the particle is exactly found at x is denoted $|x\rangle$, then $\psi(x) = \langle x \,|\, \psi\rangle$ and the probability density is $P(x) = |\psi(x)|^2$. Consider the wavefunction and probability distribution shown in Fig. 15 describing a particle in this interval. We can use the expression for information given by Eq. (13) to calculate the amount of position information which we obtain from the wavefunction. In this case that is 1.32 bits. It is somewhat more constrained than if it was localized over half the distance, which would correspond to 1 bit of information gain. The wavefunction gives us *some* information about the position, but not complete information. If our detection apparatus gave us discretized information that the particle's position was in a particular bin of width Δx, we could use Eq. (14) in a similar way. Note that we are implicitly assuming, as with any probability distribution, that there is an accessible region (AR) (see the discussion in Sect. 2) that we know the particle is in; here that is the interval $[0, L]$

Given the quantum state $\psi(x)$ we could as well ask about the results of measurements of the particle momentum p and use the same formalism to calculate the information about momentum we receive from knowing the wavefunction, I_p. This value is not the same as the information about position, and again depends on the range of momentum values considered to be the AR.

Time Dependence The operator $\hat{H}$ is the Hamiltonian operator representing the total energy of the system. For an isolated system described by a time-independent Hamiltonian H, the time development of the state vector is determined by the

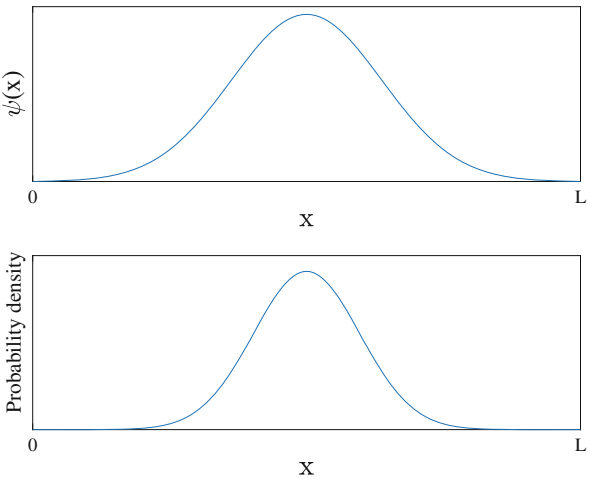

Fig. 15 Quantum wavefunction and probability density defined for a particle with a position on the continuous interval $x \in [0, L]$. The top figure shows a wavefunction $\psi(x)$ and the bottom figure the associated probability density. Given a probability density $P(x)$, one can ask how much information about the *position* of the particle does the wavefunction provide. In this case the result, from Eq. (13), is 1.32 bits. That is the Shannon measure of the information (SMI) gained by knowing the probability distribution compared to a uniform distribution over the accessible region, which is here the interval $[0, L]$

Schrödinger equation. If the state of the system at $t = 0$ is given by $|\psi(0)\rangle$, then at any future time we can determine the state by solving the differential equation

$$\boxed{i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = \hat{H} |\psi(t)\rangle} \tag{118}$$

or by evaluating the integral form

$$|\psi(t)\rangle = e^{-i\frac{\hat{H}}{\hbar}t} |\psi(0)\rangle \,. \tag{119}$$

Although the time evolution of the quantum state is deterministic, the results of measurements of the state are not, and measurements happen all the time. We lack a good account of how to describe precisely what circumstances create a measurement—the "measurement problem." But it is clear that measurements do occur, transcript of the history of the physical world is not determined by just the initial state and the physical law, but in addition by a vast number of events that could have gone one way or the other, with prescribed probabilities. One particular outcome actually occurred and was, so to speak, written into the transcript of history.

The physical law, in this case quantum mechanics, yields probabilities for possible outcomes of measurements and the Shannon measure gives us a concrete way of quantifying how much information about those outcomes the physical law

provides. In the clockwork universe of Laplace's demon, the physical law provided certainty about outcomes, given a complete description of the physical state. It turns out that the physical law simply does not do that. The physical law constrains, but does not completely constrain, the outcome of measurements. This is a profound fact about the nature of the physical law. Nor is this just a feature of the present state of quantum theory. Recent Bell test experiments confirm to astonishing accuracy that this is a feature of the physical world quite independent of quantum mechanics [15–18]. Any future successor theory would have to contain this feature—the future of the physical world is not completely constrained by the physical law. It retains some "freedom."

6.3 Open Quantum Systems and Density Operators

Pure isolated quantum states of a physical system are described by a state vector $|\psi\rangle$. Often we are dealing with a system A that is not isolated but interacting with a very large system B, which could be a thermal bath or simply the rest of the universe. We can then no longer describe the system with a state vector but must employ the formalism of the density matrix. The density matrix folds in two kinds of probability: that due to fundamental quantum indeterminacy and that due to our practical ignorance of the details of the state of a large system. It is helpful to derive it here so we can see exactly where this quantum probability and classical probability are brought together.

The starting point is writing the quantum state for the global system. We can write the state of the A system as a linear combination of basis states $|\alpha_i\rangle$. The basis states for the large system B are $|\beta_m\rangle$. The state describing the combined global system of A and B can then be written

$$|\psi\rangle = \sum_{i,m} C_{im} \, |\alpha_i ; \beta_m\rangle \tag{120}$$

where

$$\langle \alpha_i ; \beta_m \, | \, \alpha_j ; \beta_n \rangle = \delta_{ij} \delta_{mn} . \tag{121}$$

The sum over m here is over a very large number of possible states for the bath (or universe). We define the global density operator

$$\hat{\rho} = |\psi\rangle \, \langle\psi| \tag{122}$$

$$= \sum_{i,j,m,n} \underbrace{C_{im} C^*_{jn}}_{\rho_{im;\,jn}} |\alpha_i ; \beta_m\rangle \, \langle \alpha_j ; \beta_n | \tag{123}$$

$$= \sum_{i,j,m,n} |\alpha_i;\beta_m\rangle\, \rho_{im;jn} \langle\alpha_j;\beta_n| \tag{124}$$

We now focus on the target system A. Any operator $\hat{Q}^A$ which acts only on the A subsystem can be written

$$\hat{Q}^A = \sum_{i,j,m} |\alpha_i;\beta_m\rangle\, Q_{ij} \langle\alpha_j;\beta_m| . \tag{125}$$

The expectation value of Q we can write

$$\left\langle Q^A \right\rangle \equiv \left\langle \psi \,\middle|\, Q^A \,\middle|\, \psi \right\rangle = \sum_{\substack{i,\,j,\,k,\,\ell \\ m,\,n,\,p,\,q}} C^*_{kp} C_{\ell q} \langle\alpha_k;\beta_p \,|\, \alpha_i;\beta_m\rangle\, Q^A_{ij} \langle\alpha_j;\beta_m \,|\, \alpha_\ell;\beta_q\rangle . \tag{126}$$

We do the sums using (121) and exchange the symbols i and j to obtain

$$\left\langle Q^A \right\rangle = \sum_{i,j,m} C_{im} C^*_{jm} Q^A_{ji}$$

$$= \sum_{i,j} \underbrace{\left(\sum_m C_{im} C^*_{jm} \right)}_{\equiv\ \rho^A_{ij}} Q^A_{ji} \tag{127}$$

$$= \sum_{i,j} \rho^A_{ij}\ Q^A_{ji} \tag{128}$$

Therefore we can write

$$\boxed{\left\langle Q^A \right\rangle = \mathrm{Tr}(\hat{\rho}^A\ \hat{Q}^A)} \tag{129}$$

We have defined the reduced density operator ρ^A_{ij} for system A as the sum over the very large number of basis states for the environment. Our practical ignorance about the details of the large system B and its interaction with A are all hidden in this sum over the $C_{im}C^*_{jm}$ terms. Equation (129) defines the operator $\hat{\rho}^A$ as the operator on the A system which has matrix elements ρ^A_{ij} in the $|\alpha_i\rangle$ basis.

Note that comparing Eqs. (123) and (128) we have

$$\rho^A_{ij} = \sum_m \rho_{im;jm}, \tag{130}$$

which we write compactly as a partial trace of the global density matrix over the B degrees of freedom represented by the β_m states.

$$\hat{\rho}^A = \mathrm{Tr}_B(\hat{\rho}). \tag{131}$$

Several properties of the density matrix can be stated briefly. It can be easily shown that the density matrix is Hermitian and has unit trace.

$$\hat{\rho}^\dagger = \hat{\rho} \tag{132}$$

$$\mathrm{Tr}\,(\hat{\rho}) = 1 \tag{133}$$

If ρ describes a state of a system, the probability of measuring the system and finding it to be in the state ϕ is given by the expectation value of $|\phi\rangle\langle\phi|$, the projection operator for ϕ.

$$p_\phi = \mathrm{Tr}\left(|\phi\rangle\langle\phi|\ \hat{\rho}\right) \tag{134}$$

The diagonal elements of the density matrix $\rho_{i,i}$ are the probabilities of the system being found in the basis state $|\alpha_i\rangle$.

Because $\hat{\rho}$ is Hermitian it can be diagonalized by a unitary transformation. Denote the eigenvalues of $\hat{\rho}$ as ρ_ν and the eigenvectors $|\nu\rangle$. If the system is in a "pure state" that could be described by a single state vector, then all the eigenvalues will be zero except for one. If not, the state is described as "mixed." In that case the eigenvalues ρ_ν are the probability of the system being found in the state $|\nu\rangle$.

The von Neumann Entropy Von Neumann defined the quantum entropy S_{vN} to be a measure of this "mixedness."

$$\boxed{S_{vN}(\hat{\rho}) \equiv -\,\mathrm{Tr}\,(\hat{\rho}\log(\hat{\rho}))} \tag{135}$$

The von Neumann entropy is equivalent to the SMI of the eigenvalues of $\hat{\rho}$ times log(2), the conversion factor between bases for the logarithm. Alternatively, we can express the von Neumann entropy in bits, in which case it is identical to the Shannon measure of the density matrix eigenvalues.

$$S_{vN}(\hat{\rho}) = -\log(2)\ \sum_\nu \rho_\nu \log_2(\rho_\nu) \tag{136}$$

$$= \log(2)\ \mathrm{SMI}([\rho_1, \rho_2, \rho_3, \ldots, \rho_\nu, \ldots]) \tag{137}$$

$$S_{vN}^{(\mathrm{bits})}(\hat{\rho}) \equiv S_{vN}(\hat{\rho})/\log(2) \tag{138}$$

$$= \mathrm{SMI}([\rho_1, \rho_2, \rho_3, \ldots, \rho_\nu, \ldots]) \tag{139}$$

We have a set of probabilities ρ_ν and the Shannon measure tells us how much information is missing given this probability distribution. It is information about

the extent to which the system could be found to be in various of the density matrix eigenstates. Fortunately, this measure is invariant under unitary transformations, so S_{vN} is the same regardless of which basis set we use.

Time Development of the Density Matrix We can consider the case of a system that is in a mixed state, presumably because of contact with another system in the past, but which is now isolated. From the Schrödinger equation for $|\Psi\rangle$ in (123), one can derive the von Neumann equation for the time development of $\hat{\rho}$:

$$\boxed{i\hbar \frac{\partial \hat{\rho}}{\partial t} = [\hat{H}, \hat{\rho}].} \tag{140}$$

where the square brackets denote the anti-commutator. This holds of course for pure-state density matrices as well. Equation (140) is also known as the quantum Liouville equation. If the Hamiltonian is time-independent we can equivalently write the time development as a unitary transformation from the initial state directly:

$$\boxed{\hat{\rho}(t) = e^{-i\hat{H}t/\hbar}\hat{\rho}(0)e^{+i\hat{H}t/\hbar}} \tag{141}$$

The case where the system is in continous contact with the bath is, as one might suppose, much more difficult because one has to adopt some approximate model of the behavior of the bath. The most straightforward is the Lindblad formalism, but we will not explore that here.

We can say something qualitative about the evolution of the density matrix in contact with the larger system. The off-diagonal elements of the density matrix are called "coherences" because they express quantum mechanical coherence between the system A and the bath B. If the system and bath are initially in a direct product state, these coherences will quickly vanish (either exponentially or with a Gaussian shape) and the density matrix will become diagonal in the basis of energy eigenstates. The reason for this is that the system will become quantum mechanically entangled with the many degrees of freedom of the larger system and the quantum complex phases in the sum (127) will average to zero. This can be seen in moderately sized system where the global A+B system can be solved exactly [19]. If the mean interaction energy between the system and environment is E_{se}, then the coherences vanish and a time of the order of h/E_{se}.

The density matrix is the best local description we can have of the state of the system. The reality is that the system has no state—only the combined system+bath really have a quantum state. There is a loss of information from the subsystem as it interacts and entangles with the larger system, reflected by the loss of the off-diagonal elements of the density matrix. The global state is pure and the global information is undiminished, but it cannot be found in any of its subsystems. This is a uniquely quantum mechanical feature.

Statistical Mechanics for Open Quantum Systems Quantum statistical mechanics starts with the quite reasonable assumption that in thermal equilibrium all the

coherences have vanished and the density matrix is diagonal in the energy basis with probabilities given by the Boltzmann distribution [20].

$$\hat{\rho}_{\text{eq}} = \frac{e^{-\hat{H}/k_B T}}{Z} \quad \text{where} \quad Z = \text{Tr}(e^{-\hat{H}/k_B T}) \tag{142}$$

Therefore, somewhat anticlimactically, the quantum treatment of the write and erase bit operations adds little to the classical analysis in Sect. 5.3. The Hamiltonian in the basis of dot states $[|1\rangle, |2\rangle, |3\rangle]$ is given by

$$\hat{H(t)} = \sum_{k=1}^{3} |k\rangle \, E_k(t) \, \langle k| - \gamma \sum_{k=1}^{2} \Big[\, |k\rangle \, \langle k+1| + |k\rangle \, \langle k+1| \, \Big] \tag{143}$$

In order to hold the bit in the metastable state a memory requires, we want the tunneling energy γ to be small and the barrier height of E_2 in the high state to be large. If $E_1 = E_3$, we can define the barrier height as $E_b \equiv E_2 - E_1$. The effective barrier to tunneling is through this barrier is then [21]

$$\gamma_{\text{eff}} = \frac{\sqrt{E_b^2 + 8\gamma^2} - E_b}{4}. \tag{144}$$

The additional memory design requirement is then to make γ small and E_b large enough to suppress quantum tunneling for the required bit hold time. The characteristic tunneling time can be taken to be h/γ_{eff}.

With γ small there will be only slight anti-crossing of the otherwise degenerate the energy levels as E_2 moves up and down to latch or erase a bit. The Hamiltonian eigenenergies will be very close to the on-site energies of each dot E_k The energy scale of the switching is much larger than the tunneling energy by design. If the system is switched slowly enough to always keep it in the thermal ground state, which was our assumption, then the density matrix is always diagonal in the energy eigen-basis because the off-diagonal coherences are all 0. The thermodynamic entropy $S(t)$ is then identical to the von Neumann entropy $S_{vN}(t)$. As a result, each of the Figs. 9, 10, 11, 12 and 13 is essentially the same for both the classical and quantum cases. Where we could expect a difference would be if the switching speed were fast enough to drive the system out of equilibrium, or stress the ability of the system to tunnel to the equilibrium state [22].

6.4 *Non-equilibrium Quantum System: Free Expansion of an Ideal Quantum Gas*

Finally, we will look at a very-far from equilibrium situation to see the differing roles of the von Neumann entropy and the quantum entropy of outcomes. We first

extend the concept of the entropy of outcomes (111) from pure states to mixed states described by density matrices. We will now call this quantity S_Q, remembering that it is not the thermodynamic equilibrium entropy S, a state function, but rather is a dynamical measure of missing information about a particular observable. We start with a Hermitian operator $\hat{Q}$ representing an observable, its eigenvalues q_i, eigenstates $|q_i\rangle$, and the projection operator onto a particular eigenstate $|q_i\rangle\langle q_i|$. For a system described by the density matrix $\hat{\rho}$ we define:

$$p_{q_i} = \mathrm{Tr}\,(\hat{\rho}\ |q_i\rangle\langle q_i|) \tag{145}$$

$$S_Q(\hat{\rho}) \equiv -\sum_{q_i} p_{q_i} \log_2(p_{q_i}) = \mathrm{SMI}([p_{q_1}, p_{q_2}, p_{q_3}, \ldots]). \tag{146}$$

For a dynamic system we will denote $S_Q(t) = S_Q(\hat{\rho}(t))$ and see to what degree this extends the equilibrium notion of the thermodynamic entropy.

We consider a very simple system of a single-particle gas in one dimension in the low density limit. The one-particle gas models an ideal non-interacting gas in the low-density limit for which one need not take into account the Fermi-Dirac or Bose-Einstein character of the occupancy statistics. Figure 16 illustrates the physical situation. A linear array of N quantum dots, with on-site energy E_0, forms a "large" container. The spatial positions of the dots are taken to be uniform from $x_1 = 0$ to $x_N = 100$ in arbitrary units. A tunneling matrix element γ_k couples sites k and $k+1$. The Hamiltonian for the system is then

$$\hat{H} = \sum_{k=1}^{N} E_0\,|k\rangle\langle k| - \sum_{k=1}^{N-1} \gamma_k \Big[\,|k\rangle\langle k+1| + |k\rangle\langle k+1|\,\Big]. \tag{147}$$

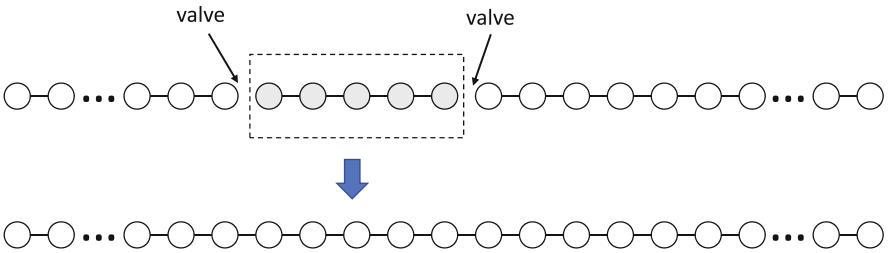

Fig. 16 Schematic view of free expansion of single-particle quantum gas on a 1D chain of quantum dots. The chain consists of a linear array of N quantum dots with near-neighbor Hamiltonian coupling energies γ. For $t < 0$ (top diagram) the particle is confined to a small initial segment of N_{init} dots in the chain. The coupling elements to the left and right of the segment are set to zero—the "valves" are closed. The initial state is in thermal equilibrium. At $t = 0$ the system is isolated from the bath and the values of coupling elements on the left and right of the segment are switched from 0 to γ (bottom diagram). The one-particle gas is therefore free to expand into the larger container. The unitary non-equilibrium evolution of the quantum density matrix is given by Eq. (141)

All the coupling elements are identical, $\gamma_k = \gamma$, except those surrounding a segment of length N_{init} dots near the center of the array. These are initially set to $\gamma_{\text{left}} = \gamma_{\text{right}} = 0$, isolating the N_{init} dots in the small container.

Before $t = 0$, the gas is in the thermal equilibrium state at temperature T and is held in the smaller container of N_{init} sites. We calculate the initial equilibrium density matrix

$$\hat{\rho}^{\text{init}} = \frac{e^{-\hat{H}_{\text{init}}/k_B T}}{\text{Tr}\left(e^{-\hat{H}_{\text{init}}/k_B T}\right)}, \tag{148}$$

where $\hat{H}_{\text{init}}$ includes only the dots in the small container.

At $t = 0$ two zero tunnelling matrix elements, γ_{left} and γ_{right} are set to the common value γ, thus opening the "valves" connecting the small container to the larger container. The initial $N_{\text{init}} \times N_{\text{init}}$ density matrix is embedded in the larger $N \times N$ density matrix describing the whole system and the time development is calculated directly from the von Neumann equation (141). The container is now assumed to be isolated from the heat bath. We calculate the case where $N_{\text{init}} = 8$, $N = 64$, and $\gamma/E_0 = 0.1$. The mean value of the energy eigenvalues is E_m and the temperature is chosen to be $T = E_m/(15k_B)$. The smaller container is offset from the center of the larger container slightly to avoid artifacts of symmetry. Time is measured in units of $\tau = \hbar/\gamma$.

The probability of dot occupancy at three snapshots in time is shown in Fig. 17. At $t = 0$ the probability is nonzero only in the smaller container with quantum confinement effects shown by the rounding of the probability at the container edges. At $t = 5\tau$ the gas is expanding freely into the surrounding container. From $t \approx 20\tau$ onward the probability fills the larger container, though because the system has no way to lose energy, quantum oscillations in the probability continue indefinitely. A snapshot at $t = 80\tau$ shows a characteristic distribution.

Figure 18 shows the probability distribution for each energy eigenstate of the system. Before the expansion ($t = 0^-$) there are $N_{\text{init}} = 8$ Hamiltonian eigenstates with the characteristic Boltzmann distribution of probabilities. The red line is the Boltzmann exponential as a continuous function of energy. Just after the valves are opened ($t = 0^+$), the number of eigenstates increases to $N = 64$. This is a far-from-equilibrium situation so the occupancy is no longer thermal. The red line again shows the Boltzmann exponential for the initial temperature, now just for comparison because there is no temperature for the system after $t = 0$. The non-equilibrium distribution in energy is perhaps surprising close to the thermal shape, though less so at low energies. The probabilities for each allowed energy do not change once the valves are open because the time evolution (141) is unitary, preserving the projection onto energy eigenstates, so the ($t = 0^+$) figure is valid for all positive times.

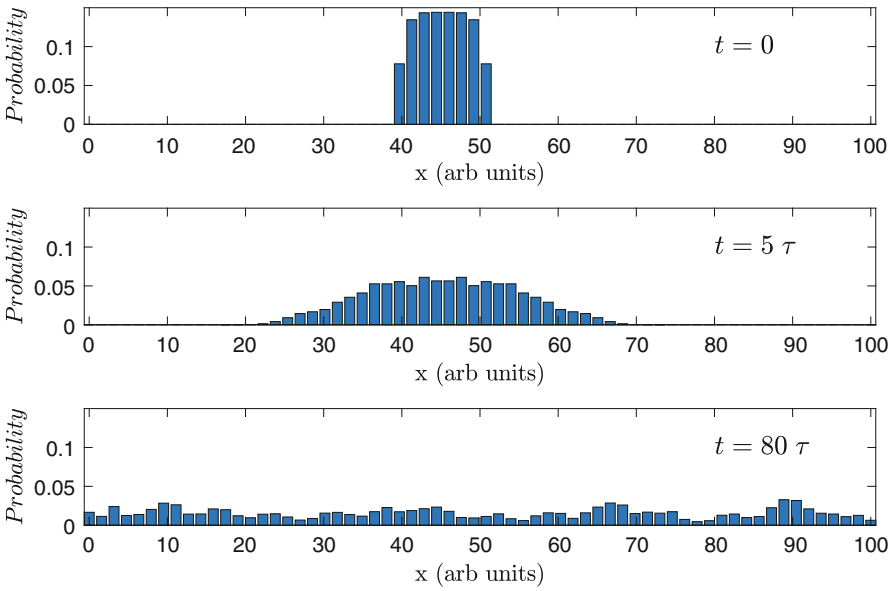

Fig. 17 Free expansion of the quantum gas in 1D. The probability density for the linear chain shown schematically in Fig. 16 is plotted at three snapshots in time during the unitary evolution of the density matrix $\rho(t)$ according to (141). At $t = 0$, the system is confined in the smaller container near the center of the array. The gas is initially at thermal equilibrium and quantum confinement effects are visible in the drop-off at the edges. The single particle gas is isolated from the thermal bath at $t = 0$ and released to expand into the surrounding larger container. Time is measured in units of $\tau = \hbar/\gamma$ where γ is the inter-dot coupling Hamiltonian matrix element (see Eq. (147). The middle plot shows the expanding gas at $t = 5\tau$. The lower plot shows the probability at $t = 80\tau$, when it has filled the container. Oscillations persist because there is no energy dissipation mechanism in the model

Figure 19 shows two measures of the entropy of the expanding gas. We again note that a thermodynamic entropy cannot be uniquely defined for this non-equilibrium situation. The dashed red line shows the von Neumann entropy $S_{vN}(t)$, measured in bits (calculated from (138)). Unsurprisingly, it is constant, 1.81 bits, throughout the expansion precisely because the free expansion is a unitary process. The eigenvalues of the density matrix do not change in time under (141).

The time development of the entropy of outcomes associated with the position operator, $S_X(t)$, calculated from (146) is shown as the solid line in Fig. 19. Initially $S_x \approx 3$, corresponding to the Shannon entropy for $8 = 2^3$ uniformly occupied dots. It is slightly lower because the probability density at $t = 0$ is not quite uniform (Fig. 17). It rises smoothly to near a value of 6, corresponding to the Shannon entropy for $64 = 2^6$ uniformly occupied dots. Again, the residual quantum oscillations visible in Fig. 17 account for the remaining difference.

In this free expansion of an ideal quantum gas, the volume was increased by a factor of 8, which would correspond to a classical increase of the thermodynamic

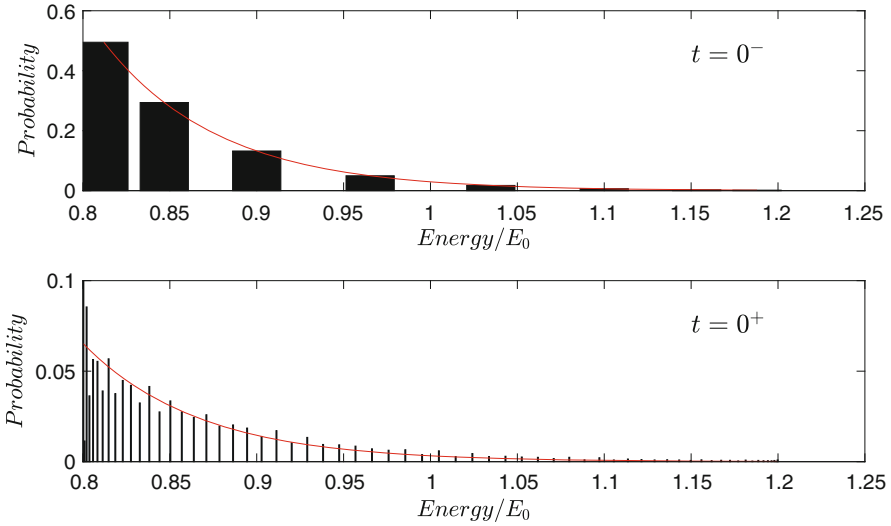

Fig. 18 The probability of each eigen-energy prior to expansion, and just after the expansion "valves" are opened for the single particle quantum gas shown in Figs. 16 and 17. Here $k_B T/E_0 = 0.067$ and $\gamma/E_0 = 0.1$. The initial state is in thermal equilibrium so the probabilities for each eigenenergy of the smaller container, with $N_{\text{init}} = 8$, are given by the Boltzmann distribution, Eq. (142) and indicated by the red line in the top figure. Immediately after the opening, there are more eigenstates because the system is now larger with $N = 64$, as shown in the lower figure. The probability associated with each eigenstate is now only approximately thermal (red line) because the system is no longer in equilibrium. The probabilities in the lower figure are constant after $t = 0^+$ throughout the unitary evolution of the isolated system

entropy by a factor of $\log_2(V_{\text{final}}/V_{\text{init}}) = 3$ (see (91)). This is nearly what we see in Fig. 19 for this entirely quantum mechanical calculation using $S_X(t)$. But it is *not* reflected in the von Neumann entropy, which is meant to capture the amount of deviation from a purity in the density matrix. The amount of "mixedness" does not change during isolated free expansion. Using $S_X(t)$ we capture what the classical version captured—the increase in the amount of *position* information that is missing. The entropy of outcomes for energy measurements S_E is constant in time with a value of 4.9 bits.

If the expanded system were to again be put in contact with the bath and allowed to come to thermal equilibrium, then we would have $S_{vN} = 4.78$ bits, $S_E = 4.88$ bits, and S_X=5.99 bits (still reflecting the small quantum edge effects). The von Neumann entropy S_{vN} is the entropy of outcomes S_Q for the case when the relevant operator is the density operator itself, $Q = \rho$, and it will always have the minimum value over the space of all Hermitian operators.

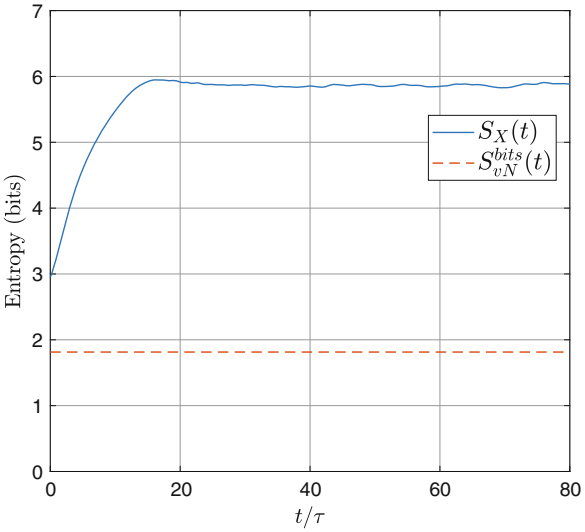

Fig. 19 Two measures of the entropy of the ideal single-particle quantum gas during free expansion. The von Neuman entropy $S_{vN}(t)$, calculated from (138, 139), is constant with a value of 1.81 bits during unitary free expansion (red dashed line). It is a measure of the lack of information about the quantum state of the system in the density matrix—its "mixedness" or deviation from purity. The entropy of outcomes for position, $S_X(t)$, defined in (146) with $\hat{Q} = \hat{X}$, is shown as the solid blue line. S_X is initially 2.96 bits. If there were no quantum size effects in the probability at $t = 0$ (Fig. 17), then we would have $S_X(0) = 3$, corresponding to equal probabilities over $8 = 2^3$ dots. As the expansion proceeds, S_X increases to a value near 6 bits, corresponding to a uniform distribution over $64 = 2^6$ dots. It is S_X that most nearly corresponds to the classical equilibrium result, given by the Sakur-Tetrode equation (91), for the entropy of a gas related to the volume it fills. Using the $S_X(t)$ as a measure of the missing position information offers a natural extension of equilibrium entropy to the non-equilibrium case

7 Discussion

Discussions of information are sometimes confused by failing to distinguish raw information from encoded information. An encoding scheme adds the mapping between physical states and logical symbols that makes a physical process, which is after all just physics in action, a meaningful logical or arithmetic process. Bits are not part of the ontology of the physical world but rather supervene on physical states.

Logically, the Shannon entropy (SMI), as a measure of missing information in a probability distribution is the most foundational concept for entropy. Probability represents a quantification of incomplete knowledge. Jaynes contribution was built on the insight that the only unbiased way to construct a probability distribution is to find that distribution which maximizes the SMI, a measure of what is unknown,

subject to the mathematical constraints that arise from what is known. This is a strict and objective procedure that is the same for all observers who have the same information available to them.

We have seen in Sect. 3 that applying the Jaynes maximum entropy principle yields mathematical results which reproduce standard statistical mechanics. Classical statistical mechanics is, after all, unnecessary if all the relevant dynamical quantities are known—that would require only mechanics. Statistical mechanics concerns what we can know about a system when the details of its dynamics are unknown. When applied to the problem of a target system in thermal equilibrium with a large heat bath, the Jaynes procedure precisely connects the information theoretic Shannon entropy with the thermodynamic entropy of von Neumann, Gibbs, Boltzmann, and Clausius. The thermodynamic entropy in equilibrium is (within a constant) the Shannon entropy of the unique probability distribution which maximizes the SMI subject to the appropriate constraints. The equilibrium thermodynamic entropy is a special case of the Shannon entropy applied to particular systems that have a well-characterized average energy, number of particles, or other macroscopic constraints.

The Jaynes formulation of physical entropy as a measure of missing information is superior to the often-encountered notion of entropy as a measure of disorder. Disorder is not a quantifiable idea—it is irredeemably subjective, completely in the mind of the beholder. Information (in the information theoretic sense measured in bits) is quantifiable and precise. The information that is missing from a given probability distribution is quantifiable and objective.

The Landauer Principle connects an information theoretic process, bit erasure, with a physical process, heat dissipation. Because it concerns a lower bound for heat dissipation we looked quantitatively at a minimal physical system with a specific encoding scheme. A key step here was to use the Jaynes definition of thermodynamic entropy to describe a memory storage device which is ipso facto not in an equilibrium state. This straightforward extension permits a quantitative analysis of the minimal thermodynamic system when a known or unknown bit is erased. We see precisely the expected heat dissipation of $k_B \log(2)$ when an unknown bit is erased, and no lower bound for the heat dissipated when erasing a known bit (with a copy preserved).

In the quantum mechanical case, the von Neumann entropy is the SMI of the eigenvalues of the density matrix. This is a measure of quantum state purity; a pure state has von Neumann entropy of 0. A quantum treatment of the minimal memory system acts essentially the same as the classical system because quantum coherences vanish in thermal equilibrium. It must be emphasized that in quantum mechanics what is unknown includes the fundamental indeterminacy of the physical law. This is now known to be not a peculiarity of the quantum mechanical description, but rather a feature of the nature of the physical world.

Grounding the entropy concept in the Shannon measure also naturally focuses attention on the less well-known quantum entropy of outcomes S_Q for measurements of an observable Q. We have seen that in the case of the free expansion of a classical gas, the quantum analogue of the classical entropy was not the von

Neumann entropy, but the entropy of outcomes for the position operator, S_X (146). This entropy is not basis-independent—it depends specifically on the observable in which one is interested. Whereas the von Neumann entropy captures the amount of information missing about the pure state of the system due to entanglement with the environment, S_Q capture the amount of missing information about measurements of the observable Q. It is applicable to both pure and mixed states.

Acknowledgements Thanks to Arieh Ben-Naim for highlighting the Jaynes approach in references [4] and [5], and for helpful conversations. Thanks also to Neal Anderson and Ken Sauer for many stimulating talks on information and related topics.

References

1. W. Porod, R. Grondin, D. Ferry, G. Porod, Phys. Rev. Lett. **52**, 232 (1984)
2. J.D. Norton, Stud. Hist. Philos. Mod. Phys. **36B**(2), 375 (2005)
3. E. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003)
4. A. Ben-Naim, *Modern Thermodynamics* (World Scientific, Singapore, 2016)
5. A. Ben-Naim, *A Farewell to Entropy: Statistical Mechanics Based on Information* (World Scientific, Singapore, 2008)
6. T. Cover, J. Thomas, *Elements of Information Theory*, 2nd edn. (Wiley, Hoboken, 2006)
7. M. Trebus, *Thermostatics and Thermodynamics* (D. Van Nostrand, New York, 1961)
8. G.M. Jochen Gemmer, M. Michel, *Quantum Thermodynamics: Emergence of Thermodynamic Behavior Within Composite Quantum Systems*, *Lecture Notes in Physics*, vol. 784 (Springer, Berlin, 2009)
9. I. Amlani, A.O. Orlov, G.L. Snider, C.S. Lent, G.H. Bernstein, Appl. Phys. Lett. **72**(17), 2179 (1998)
10. J. Christie, R. Forrest, S. Corcelli, N. Wasio, R. Quardokus, R. Brown, S. Kandel, C. Lent, Y. Lu, K. Henderson, Angew. Chem. **127**, 15668 (2015)
11. A.O. Orlov, C.S. Lent, C.C. Thorpe, G.P. Boechler, G.L. Snider, Jpn. J. Appl. Phys. **51**, 06FE10m (2012)
12. R.K. Kummamuru, A.O. Orlov, R. Ramasubramaniam, C.S. Lent, G.H. Bernstein, G.L. Snider, IEEE Trans. Electron Devices **50**(9), 1906 (2003)
13. L. Szilard, **5**, 840 (1929)
14. G. Paneru, D.Y. Lee, T. Tlusty, H.K. Pak, Phys. Rev. Lett. **120**, 020601 (2018). https://doi.org/10.1103/PhysRevLett.120.020601. https://link.aps.org/doi/10.1103/PhysRevLett.120.020601
15. M. Giustina, M.A.M. Versteegh, S. Wengerowsky, J. Handsteiner, A. Hochrainer, K. Phelan, F. Steinlechner, J. Kofler, J.A. Larsson, C. Abellán, W. Amaya, V. Pruneri, M.W. Mitchell, J. Beyer, T. Gerrits, A.E. Lita, L.K. Shalm, S.W. Nam, T. Scheidl, R. Ursin, B. Wittmann, A. Zeilinger, Phys. Rev. Lett. **115**, 250401 (2015). https://doi.org/10.1103/PhysRevLett.115.250401. http://link.aps.org/doi/10.1103/PhysRevLett.115.250401
16. J. Handsteiner, A.S. Friedman, D. Rauch, J. Gallicchio, B. Liu, H. Hosp, J. Kofler, D. Bricher, M. Fink, C. Leung, A. Mark, H.T. Nguyen, I. Sanders, F. Steinlechner, R. Ursin, S. Wengerowsky, A.H. Guth, D.I. Kaiser, T. Scheidl, A. Zeilinger, Phys. Rev. Lett. **118**, 060401 (2017). https://doi.org/10.1103/PhysRevLett.118.060401. https://link.aps.org/doi/10.1103/PhysRevLett.118.060401

17. B. Hensen, H. Bernien, A.E. Dreau, A. Reiserer, N. Kalb, M.S. Blok, J. Ruitenberg, R.F.L. Vermeulen, R.N. Schouten, C. Abellan, W. Amaya, V. Pruneri, M.W. Mitchell, M. Markham, D.J. Twitchen, D. Elkouss, S. Wehner, T.H. Taminiau, R. Hanson, Nature **526**, 682 (2015). https://doi.org/10.1038/nature15759
18. L.K. Shalm, E. Meyer-Scott, B.G. Christensen, P. Bierhorst, M.A. Wayne, M.J. Stevens, T. Gerrits, S. Glancy, D.R. Hamel, M.S. Allman, K.J. Coakley, S.D. Dyer, C. Hodge, A.E. Lita, V.B. Verma, C. Lambrocco, E. Tortorici, A.L. Migdall, Y. Zhang, D.R. Kumor, W.H. Farr, F. Marsili, M.D. Shaw, J.A. Stern, C. Abellán, W. Amaya, V. Pruneri, T. Jennewein, M.W. Mitchell, P.G. Kwiat, J.C. Bienfang, R.P. Mirin, E. Knill, S.W. Nam, Phys. Rev. Lett. **115**, 250402 (2015). https://doi.org/10.1103/PhysRevLett.115.250402. https://link.aps.org/doi/10.1103/PhysRevLett.115.250402
19. E. Blair, C. Lent, J. Appl. Phys. **113**, 124302 (2013)
20. R.P. Feynman, *Statistical Mechanics: A Set of Lectures* (W.A. Benjamin, Reading, 1972)
21. J.P. Timler, Energy dissipation and power gain in quantum-dot cellular automata, Ph.D. thesis, University of Notre Dame, 2003
22. E.P. Blair, S.A. Corcelli, C.S. Lent, J. Chem. Phys. **145**(1), 014307 (2016)

Conditional Erasure and the Landauer Limit

Neal G. Anderson

Contents

1 Introduction

Landauer's Principle (LP), in its original and most widely recognized form, stipulates that the erasure of information from a physical system necessarily dissipates heat into the system's environment [1]. The associated lower bound on the dissipative cost of erasure—$k_B T \ln(2)$ of environmental heating per erased bit of information—is widely known as the Landauer limit. Related work of Bennett [2, 3] associated the Landauer limit only with irreversible erasure processes, and

N. G. Anderson (✉)
Department of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst, MA, USA
e-mail: anderson@ecs.umass.edu

C. S. Lent et al. (eds.), *Energy Limits in Computation*,
https://doi.org/10.1007/978-3-319-93458-7_2

suggested that there is no such lower limit on dissipation for reversible erasure processes that utilize a durable record of the system's pre-erasure state. We will call this lower bound on the dissipative cost of reversible erasure—*zero* environmental heating per erased bit of information—the Landauer-Bennett limit.

Both the Landauer and Landauer-Bennett limits have been discussed and analyzed for decades, primarily through theoretical analyses of idealized classical and quantum-mechanical model systems, and both have long been controversial [3–25]. Some of the controversy can be traced to flaws in the debate, e.g. imprecise or incomplete definition of information and other key quantities; imprecise or incomplete specification of the operations, processes, system boundaries, and physical scenarios involved; excessive reliance on simple model systems and over-generalization of results obtained from analyses of these systems; and conflation of inviolable (if possibly unachievable) bounds with practically achievable limits. Much of the debate is, however, rooted in substantive methodological and interpretive concerns, including the proper application of thermodynamics, the connections and/or distinctions between physical and information-theoretic entropies, and the connections/distinctions between physical and logical reversibility. In any case, decades-old controversies over the Landauer and Landauer-Bennett limits live on as experimentalists begin to probe these limits experimentally [26–29] and as their potential implications for the future of computation attract renewed interest [30, 31].

This state of affairs motivates rigorous proof of very general bounds on the dissipative cost of information erasure—both for erasure protocols that invoke records of pre-erasure system states and those that do not—within a common theoretical framework. This framework should capture the essential physical distinction between conditional and unconditional erasure protocols, root the associated dissipative costs in fundamental physical law (as opposed to the details of particular realizations), distinguish information from physical entropy, and withstand the methodological objections commonly raised in the literature.

Such proofs are provided in this work. Using a very general and ecumenical physical description of Landauer erasure, and assuming little more than distinguishability of encoding states in data-bearing systems and the validity of non-relativistic quantum theory, we obtain dissipation bounds for conditional and unconditional Landauer erasure of information from quantum dynamics and entropic inequalities alone. Shannon entropy, which most commonly plays the role of an information measure in expressions of Landauer-like bounds, emerges naturally in our proof of the unconditional erasure bound: it is neither inserted "by hand" at the outset nor assumed to be equivalent to thermodynamic entropy or any other physical entropy germane to the problem. The underlying theoretical methodology is, we argue, immune to common objections to (mostly thermodynamic) demonstrations, arguments, and proofs of the Landauer and Landauer-Bennett limits, expressed most precisely and forcefully by John D. Norton [16], and is consistent with results from recent experimental probes. Our results are further supported by bounds obtained from a much more general physical-information-theoretic approach, which uses an overtly physical information measure that differs fundamentally from Shannon's mathematical entropy but that coincides with it for the case of distinguishable encoding states. Our analysis thus provides fundamental support for interpretation

of Landauer's limit as a lower bound on environmental heating *contributed by* information loss in unconditional Landauer erasure, and for the absence of any such dissipative contributions in conditional Landauer erasure.

The remainder of this chapter is organized as follows. In Sect. 2 we define Landauer erasure in terms of essential properties of the physical state transformations required to reset a physical system's state to a "standard state," and define the corresponding erasure cost as the initial-state-averaged system-to-environment energy transfer resulting from such state transformations. In Sect. 3, we obtain a very general bound on the cost of erasure so defined, and identify it as a generalization of the Landauer-Bennett bound for conditional erasure (Sect. 3.1). We then specialize this proof to preclude conditional erasure operations—operations that are "preselected" to match the initial system state—and identify the resulting bound as a generalization of the Landauer limit (Sect. 3.2). Next, we discuss the physical origin of the Landauer limit and the role of conditioning within our descriptive framework, and elaborate on the subtle role that delineation of the "erasable system" boundary plays in erasure protocol classification (Sect. 3.3). In Sect. 4, we catalog some recent methodological objections to existing proofs of the Landauer and Landauer-Bennett limits and show how they are sidestepped in our framework. In Sect. 5, we remark on the consistency of recent experimental results with the results of our work. We conclude in Sect. 6, and place the bounds obtained here in the context of the more general and thoroughly physical approach we have described elsewhere.

2 Landauer Erasure

2.1 Physical Description

We begin with a very general physical description of Landauer erasure, i.e. the resetting of an information-bearing physical system to a standard state. Consider a system $\mathcal{S}$ that can be initially prepared in any of N *distinguishable* states $\hat{\rho}_i^{\mathcal{S}}$ drawn from a set $\{\hat{\rho}_i^{\mathcal{S}}\} = \{\hat{\rho}_0^{\mathcal{S}} ... \hat{\rho}_{N-1}^{\mathcal{S}}\}$, and can thus unambiguously encode classical information expressed in an N-ary alphabet. We define a *Landauer erasure protocol* [1] as an operation or sequence of operations on $\mathcal{S}$ that, for every initial state $\hat{\rho}_i^{\mathcal{S}}$, leaves the system $\mathcal{S}$ in a standard reset state $\hat{\rho}_{\text{reset}}^{\mathcal{S}}$ (which may or may not coincide with one of the $\hat{\rho}_i^{\mathcal{S}}$). We assume that $\mathcal{S}$ couples to an environment $\mathcal{E}$ throughout the erasure process, and that the environment is *initially* in a thermal state $\hat{\rho}_{\text{th}}^{\mathcal{E}}$ at temperature T. With this, Landauer erasure can be formally defined by the set

$$\{\rho_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}} \to \hat{\rho}_i^{\mathcal{S}\mathcal{E}'}\}$$

of N state transformations of $\mathcal{S}\mathcal{E}$, for which the final local state of $\mathcal{S}$ is

$$Tr_{\mathcal{E}}[\hat{\rho}_i^{\mathcal{S}\mathcal{E}'}] = \hat{\rho}_{\text{reset}}^{\mathcal{S}} \quad \forall i.$$

Global closure of $\mathcal{SE}$ implies unitary evolution of the system-environment composite during Landauer erasure, consistent with joint evolution of $\mathcal{SE}$ governed by the time-dependent Schrodinger equation. This is to say that, for each of the N $\hat{\rho}_i^{\mathcal{S}} \in \{\hat{\rho}_i^{\mathcal{S}}\}$, there exists a unitary operator $\hat{U}_i \in \{\hat{U}_i\}$ that transforms the state of $\mathcal{SE}$ as

$$\hat{U}_i(\rho_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}})\hat{U}_i^{\dagger} = \hat{\rho}_i^{\mathcal{SE}'}$$

such that $Tr_{\mathcal{E}}[\hat{\rho}_i^{\mathcal{SE}'}] = \hat{\rho}_{\text{reset}}^{\mathcal{S}}$, i.e. so the required state reset is achieved. Here

$$\hat{U}_i = \mathcal{T}\left\{\exp\left[-\frac{i}{\hbar}\int_0^{t_{\text{er}}} \hat{H}_i^{\mathcal{SE}}(t')dt'\right]\right\}$$

with

$$\hat{H}_i^{\mathcal{SE}}(t) = \left(\hat{H}_{\text{self}}^{\mathcal{S}} + \hat{H}_{\text{self}}^{\mathcal{E}} + \hat{H}_{\text{int}}^{\mathcal{SE}} + V_i^{\mathcal{S}}(t)\right)$$

where t_{er} is the duration of the erasure operation (initiated at $t = 0$), $\hat{H}_{\text{self}}^{\mathcal{S}}$ and $\hat{H}_{\text{self}}^{\mathcal{E}}$ are the system and environment self Hamiltonians, $\hat{H}_{\text{int}}^{\mathcal{SE}}$ is the system–environment interaction Hamiltonian, $V_i^{\mathcal{S}}(t)$ is the time-dependent potential applied to the system to reset the i-th system state $\hat{\rho}_i^{\mathcal{S}}$, and $\mathcal{T}$ is the Dyson time-ordering operator. We note for later reference that the Hamiltonian has been defined so all differences between the various $\hat{U}_i$ stem exclusively from differences in the applied potential $V_i^{\mathcal{S}}(t)$.

This is a very general and "ecumenical" physical description of Landauer erasure. It is based on a high-level specification of requirements that states and state transformations must meet to (1) correspond to Landauer erasure and (2) obey physical law as expressed in the postulates of quantum mechanics. This is to be contrasted with low-level descriptions—explicit physical models—of erasure processes in particular realizations of information-bearing systems. Our high-level description aims to capture only those essential features of both conditional erasure protocols that employ operations conditioned on the initial state of $\mathcal{S}$—or, equivalently, on a physical copy of the state of $\mathcal{S}$—and unconditional erasure protocols that involve no such conditioning. The distinction between conditional and unconditional protocols is conceptually significant and provably consequential for the dissipative cost of erasure, as shown below.

2.2 Energy Cost of Erasure

Our objective is to lower bound the average environment energy increase

$$\langle \Delta\langle E_i^{\mathcal{E}}\rangle\rangle \equiv \sum_i p_i \Delta\langle E_i^{\mathcal{E}}\rangle \tag{1}$$

for Landauer erasure protocols. The average is relevant to a large collection $\{p_i, \hat{\rho}_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}}\}$ of system-environment composites, a fraction p_i of which have the system $\mathcal{S}$ initially prepared in the state $\hat{\rho}_i^{\mathcal{S}}$ and all of which are in contact with thermal environments at temperature T. (It applies equally well to a large number of state resets performed on a single system-environment composite, with $\mathcal{SE}$ initially prepared in the joint state $\hat{\rho}_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}}$ in a fraction p_i of erasure trials.) Here $\Delta\langle E_i^{\mathcal{E}}\rangle$ is the increase in the *expected value* of the environment energy associated with resetting the initial system states $\hat{\rho}_i^{\mathcal{S}}$ to the standard state $\hat{\rho}_{\text{reset}}^{\mathcal{S}}$.

For later reference, we write Eq. (1) out as

$$\langle \Delta\langle E_i^{\mathcal{E}}\rangle\rangle = \sum_i p_i \left(\langle E_i^{\mathcal{E}'}\rangle - \langle E_i^{\mathcal{E}}\rangle\right)$$

$$= \left(\sum_i p_i Tr_{\mathcal{E}}\left[\hat{\rho}_i^{\mathcal{E}'} \hat{H}_{\text{self}}^{\mathcal{E}}\right]\right) - \langle E_{\text{th}}^{\mathcal{E}}\rangle$$

where $\langle E_{\text{th}}^{\mathcal{E}}\rangle = Tr_{\mathcal{E}}\left[\hat{\rho}_{\text{th}}^{\mathcal{E}} \hat{H}_{\text{self}}^{\mathcal{E}}\right]$. Noting that the final environment states are of the form

$$\hat{\rho}_i^{\mathcal{E}'} = Tr_{\mathcal{S}}\left[\hat{\rho}_i^{\mathcal{SE}'}\right] = Tr_{\mathcal{S}}\left[\hat{U}_i \hat{\rho}_i^{\mathcal{SE}} \hat{U}_i^{\dagger}\right]$$

we have

$$\langle \Delta\langle E_i^{\mathcal{E}}\rangle\rangle = Tr_{\mathcal{E}}\left[\sum_i p_i Tr_{\mathcal{S}}\left[\hat{U}_i \hat{\rho}_i^{\mathcal{SE}} \hat{U}_i^{\dagger}\right] \hat{H}_{\text{self}}^{\mathcal{E}}\right] - \langle E_{\text{th}}^{\mathcal{E}}\rangle$$

or

$$\langle \Delta\langle E_i^{\mathcal{E}}\rangle\rangle = Tr_{\mathcal{E}}\left[Tr_{\mathcal{S}}\left[\sum_i p_i \left(\hat{U}_i \hat{\rho}_i^{\mathcal{SE}} \hat{U}_i^{\dagger}\right)\right] \hat{H}_{\text{self}}^{\mathcal{E}}\right] - \langle E_{\text{th}}^{\mathcal{E}}\rangle \tag{2}$$

where the linearity of the partial trace operation $\left(\sum_i p_i Tr_{\mathcal{S}}\left[\hat{\rho}_i^{\mathcal{SE}}\right] = Tr_{\mathcal{S}}\left[\sum_i p_i \hat{\rho}_i^{\mathcal{SE}}\right]\right)$ has been used in the final equality.

3 Dissipation Bounds for Landauer Erasure

3.1 Conditional Erasure Bound

We now obtain general lower bounds on the energy cost $\langle \Delta\langle E_i^{\mathcal{E}}\rangle\rangle$ for Landauer erasure, as straightforwardly defined in Eq. (1), from very fundamental considerations.

We begin by noting that, for the initial and final states of $\mathcal{SE}$,

$$S(\hat{\rho}_i^{\mathcal{SE}'}) \leq S(\hat{\rho}_{\text{reset}}^{\mathcal{S}}) + S(\hat{\rho}_i^{\mathcal{E}'})$$

$$S(\rho_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}}) = S(\hat{\rho}_i^{\mathcal{S}}) + S(\hat{\rho}_{\text{th}}^{\mathcal{E}}).$$

The first follows from the subadditivity of von Neumann entropy for general joint states of bipartite composite systems and the second for the additivity of this entropy for separable joint states.[1] Since unitary evolution preserves von Neumann entropy, we can equate the initial state entropy $S(\rho_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}})$ and the final state entropy $S(\hat{\rho}_i^{\mathcal{SE}'})$ of $\mathcal{SE}$ and rearrange to get

$$S(\hat{\rho}_i^{\mathcal{E}'}) - S(\hat{\rho}_{\text{th}}^{\mathcal{E}}) \geq -[S(\hat{\rho}_{\text{reset}}^{\mathcal{S}}) - S(\hat{\rho}_i^{\mathcal{S}})]$$

or

$$\Delta S_i^{\mathcal{E}} \geq -\Delta S_i^{\mathcal{S}}$$

where $\Delta S_i^{\mathcal{S}} = S(\hat{\rho}_{\text{reset}}^{\mathcal{S}}) - S(\hat{\rho}_i^{\mathcal{S}})$ and $\Delta S_i^{\mathcal{E}} = S(\hat{\rho}_i^{\mathcal{E}'}) - S(\hat{\rho}_{\text{th}}^{\mathcal{E}})$. Applying Partovi's inequality [32] (see Appendix), we have for the i-th preparation $\rho_i^{\mathcal{SE}} = \rho_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}}$ and unitary $\hat{U}_i$,

$$\Delta\langle E_i^{\mathcal{E}}\rangle \geq -k_B T \ln(2) \Delta S_i^{\mathcal{S}}.$$

From this, and from Eq. (1), the average environmental energy increase is lower bounded simply as

$$\langle \Delta\langle E_i^{\mathcal{E}}\rangle\rangle \geq -k_B T \ln(2) \sum_i p_i \Delta S_i^{\mathcal{S}}. \tag{3}$$

This bound is proportional to the average decrease in self entropy of the individual system states $\hat{\rho}_i^{\mathcal{S}}$ as they are transformed to the reset state. Conspicuously absent is any term explicitly related to information loss in the erasure process.

The bound (3) applies to *conditional* Landauer erasure protocols, i.e. protocols for which the operations $\{\hat{U}_i\}$ applied to the system to achieve erasure are properly matched to—are conditioned upon—initial system states $\hat{\rho}_i^{\mathcal{S}}$. This bound presumes existence of a set $\{\hat{U}_i\}$ of N *generally different* unitary evolution operators—a set of N time-dependent applied potentials $V_i^{\mathcal{S}}(t)$—that guide transformation of each of the N pre-erasure states $\hat{\rho}_i^{\mathcal{S}} \rightarrow \hat{\rho}_{\text{reset}}^{\mathcal{S}}$ to the reset state. This is to say that selection of the applied potential in any trial is conditioned upon the pre-erasure state of $\mathcal{S}$, so the appropriate evolution operator $\hat{U}_i$ is selected and applied only when the pre-erasure state of the system is $\hat{\rho}_i^{\mathcal{S}}$.

[1] See the Appendix, where established properties of von Neumann entropy, unitary transformations, and trace operations used in this work are cataloged for convenience.

The tailored application of selected operations required for conditional erasure can be achieved in a variety of ways. However, any conditional erasure protocol presumes the existence or creation of a record—*external* to $\mathcal{S}$—of the pre-erasure state of $\mathcal{S}$. The appropriate operation can be selected and applied by an agent with prior knowledge (a remembered record) of the initial system state, by an automatous operation that can read an existing physical record of the initial state of $\mathcal{S}$ (or create one if necessary[2]), or automatically through interaction of $\mathcal{S}$ with another system whose state holds a record or copy of the pre-erasure state of $\mathcal{S}$. The bound (3) generally applies to all such conditional Landauer erasure protocols, which we hereafter call ERASE WITH COPY protocols.[3]

Note that the bound (3) is simply $\Delta\langle E_i^{\mathcal{E}}\rangle \geq 0$ if, as in those cases most widely considered in the literature, all initial states have the same entropy as the reset state. This corresponds to the Landauer-Bennett limit, reflecting no lower bound on the cost of information erasure for ERASE WITH COPY operations. In the general case (3), $\langle\Delta\langle E_i^{\mathcal{E}}\rangle\rangle$ can be positive, negative, or zero as a result of entropy flows that are not related to information erasure and that may involve only reversible system-environment energy transfer.

3.2 Unconditional Erasure Bound and the Landauer Limit

We next consider unconditional Landauer erasure, in which resetting of a system to a standard state is achieved by application of a fixed operation independent of the pre-erasure state of the system. Such protocols are more common in practice (e.g. CLEAR and RESET operations in standard digital hardware), and are familiar from thought experiments long discussed in the literature. The two most familiar of these classic thought experiments are (1) unconditional erasure of a bit encoded in a Szilard-engine by removing the partition and pushing the particle to a "standard" side of the chamber via isothermal compression, and (2) erasure of a one-bit double-potential-well memory by a time-dependent variation of the potential that lowers the

[2]An agent or automaton without an existing record of the pre-erasure system state can create one through measurement as part of a more complex conditional erasure protocol, since the N pre-erasure states are mutually orthogonal and thus distinguishable from one another. Formal treatment of any such protocol would, however, have to include an additional subsystem that registers measurement outcomes and account for physical costs associated with the measurement process (e.g. creation of system/apparatus correlations and/or erasure or overwriting of the measurement outcomes on each use). This is discussed further in Sect. 4.

[3]We adopt Orlov's [29] labeling of erasure protocols that do and do not utilize records of pre-erasure system states as ERASE WITH COPY and ERASE WITHOUT COPY protocols, respectively. Here, WITH and WITHOUT will generally mean "conditioned upon" and "not conditioned upon," so WITH COPY means both "in the presence of" and "with active use of" a copy or record. Note the physical state serving as an external record need not be an identical copy of the pre-erasure state of $\mathcal{S}$; it need only provide an unambiguous physical identifier of which encoding state $\hat{\rho}_i^{\mathcal{S}}$ the system $\mathcal{S}$ is in prior to erasure.

barrier between the wells, pushes the particle to a standard side of the well, and then restores the barrier.

The essential feature of unconditional erasure protocols—the feature that distinguishes them from the conditional protocols discussed in Sect. 3.1—is this: in unconditional protocols, *all* N state transformations

$$\rho_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}} \rightarrow \hat{\rho}_i^{\mathcal{S}\mathcal{E}'} : Tr_{\mathcal{E}}[\hat{\rho}_i^{\mathcal{S}\mathcal{E}'}] = \hat{\rho}_{\text{reset}}^{\mathcal{S}}$$

required for Landauer erasure are achieved by blind application of a *single* unitary operation $\hat{U}$ to $\mathcal{S}\mathcal{E}$ as

$$\hat{U}(\rho_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}})\hat{U}^{\dagger} = \hat{\rho}^{\mathcal{S}\mathcal{E}'}.$$

Unconditional erasure requires a universal unitary $\hat{U}$—a single time-dependent applied potential $V^{\mathcal{S}}(t)$—that is up to this task.

To lower bound the energy cost of unconditional Landauer erasure, note that it is a special case of conditional erasure with $\hat{U}_i = \hat{U}$ for all i. We can thus specialize the expression (2) for generally unconditional erasure accordingly, by substituting $\hat{U}_i \rightarrow \hat{U}$ for all i, to obtain

$$\langle \Delta \langle E_i^{\mathcal{E}} \rangle \rangle = Tr_{\mathcal{E}} \left[Tr_{\mathcal{S}} \left[\sum_i p_i \left(\hat{U} \hat{\rho}_i^{\mathcal{S}\mathcal{E}} \hat{U}^{\dagger} \right) \right] \hat{H}^{\mathcal{E}} \right] - \langle E_{\text{th}}^{\mathcal{E}} \rangle. \tag{4}$$

Because the time-evolution operator is now independent of the pre-erasure state, the linearity of unitary similarity transformations can be applied to obtain

$$\langle \Delta \langle E_i^{\mathcal{E}} \rangle \rangle = Tr_{\mathcal{E}} \left[Tr_{\mathcal{S}} \left[\hat{U} \left(\sum_i p_i \hat{\rho}_i^{\mathcal{S}\mathcal{E}} \right) \hat{U}^{\dagger} \right] \hat{H}^{\mathcal{E}} \right] - \langle E_{\text{th}}^{\mathcal{E}} \rangle. \tag{5}$$

For initial states of the form $\hat{\rho}_i^{\mathcal{S}\mathcal{E}} = \hat{\rho}_i^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}}$, this is

$$\langle \Delta \langle E_i^{\mathcal{E}} \rangle \rangle = Tr_{\mathcal{E}} \left[Tr_{\mathcal{S}} [\hat{U} \left(\hat{\rho}^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}} \right) \hat{U}^{\dagger}] \hat{H}^{\mathcal{E}} \right] - \langle E_{\text{th}}^{\mathcal{E}} \rangle \tag{6}$$

with

$$\hat{\rho}^{\mathcal{S}} = \sum_i p_i \hat{\rho}_i^{\mathcal{S}}. \tag{7}$$

The expression (6), which is equivalent to the straightforward average (2) over the N individual processes, is *equivalent to* the energy increase of the environment for the *single* state transformation

$$\hat{\rho}^{\mathcal{S}} \otimes \hat{\rho}_{\text{th}}^{\mathcal{E}} \rightarrow \hat{\rho}^{\mathcal{S}\mathcal{E}'} \tag{8}$$

of $\mathcal{SE}$ (via $\hat{U}$), with the initial state of $\mathcal{S}$ described by the density operator $\hat{\rho}^{\mathcal{S}}$ given by (7) (for which $Tr_{\mathcal{E}}[\hat{\rho}^{\mathcal{SE}'}] = \hat{\rho}^{\mathcal{S}}_{\text{reset}}$). Because of this mathematical equivalence, the desired unconditional erasure bound can be obtained by lower bounding the environment energy increase via Partovi's inequality for the "surrogate" state transformation (8), which yields

$$\langle \Delta \langle E_i^{\mathcal{E}} \rangle \rangle \geq -k_B T \ln(2) \left[S(\hat{\rho}^{\mathcal{S}}) - S(\hat{\rho}^{\mathcal{S}}_{\text{th}}) \right].$$

It follows from the mutual orthogonality of the $\hat{\rho}_i^{\mathcal{S}}$ that

$$S(\hat{\rho}^{\mathcal{S}}) = S\left(\sum_i p_i \hat{\rho}_i^{\mathcal{S}} \right) = \sum_i p_i S(\hat{\rho}_i^{\mathcal{S}}) + H(\{p_i\}) \tag{9}$$

where

$$H(\{p_i\}) = -\sum_i p_i \log_2 p_i \tag{10}$$

is the Shannon entropy [33] of the pmf $\{p_i\}$ (sometimes called the "preparation entropy" or "encoding entropy"). This finally yields the bound

$$\langle \Delta \langle E_i^{\mathcal{E}} \rangle \rangle \geq -k_B T \ln(2) \sum_i p_i \left[\Delta S_i^{\mathcal{S}} + H(\{p_i\}) \right]$$

or

$$\langle \Delta \langle E_i^{\mathcal{E}} \rangle \rangle \geq -k_B T \ln(2) \sum_i p_i \Delta S_i^{\mathcal{S}} - k_B T \ln(2) H(\{p_i\}) \tag{11}$$

for unconditional erasure.

The Shannon entropy $H(\{p_i\})$, which emerged in derivation of the bound (11), is commonly taken as a measure of the information initially encoded in $\mathcal{S}$—and subsequently erased from $\mathcal{S}$—in such scenarios, i.e. when the i-th symbol of a classical source alphabet is physically encoded in $\mathcal{S}$ by preparing $\mathcal{S}$ in the state $\hat{\rho}_i^{\mathcal{S}}$. If we adopt this measure, and write $H(\{p_i\}) = -\Delta I_{\text{er}}$ since $H(\{p_i\})$ of information is *lost* from $\mathcal{S}$ in erasure, we have

$$\langle \Delta \langle E_i^{\mathcal{E}} \rangle \rangle \geq -k_B T \ln(2) \sum_i p_i \Delta S_i^{\mathcal{S}} + k_B T \ln(2) \Delta I_{\text{er}} \tag{12}$$

where, again, $\Delta S_i^{\mathcal{S}} = S(\hat{\rho}_i^{\mathcal{S}'}) - S(\hat{\rho}^{\mathcal{S}}_{\text{th}})$.

The first term in the unconditional erasure bound (12), which reflects the average entropy change of the encoding states incurred in resetting, is equivalent to the conditional erasure bound (3). This term can, again, be positive, negative, or zero,

and is not specifically related to information erasure. The second term, which is absent from the conditional erasure bound (3), reflects a contribution to the energy cost of Landauer erasure attributable to information erasure. This term is nonnegative and represents an irreversible transfer of energy to the environment.

If all initial states $\hat{\rho}_i^{\mathcal{S}}$ have the same entropy as the reset state—again the canonical case considered in the literature—the bound (12) takes the familiar form

$$\langle \Delta \langle E_i^{\mathcal{E}} \rangle \rangle \geq k_B T \ln(2) \Delta I_{\text{er}}.$$

This bound, which has the standard form of Landauer's limit, says that unconditional erasure of information from a physical system—if quantified by the Shannon entropy of the encoding—*contributes* at least $k_B T \ln(2)$ of energy per erased bit of information to the system's (thermal) environment.

It is entirely reasonable to expect that, as the above bounds suggest, unconditional Landauer erasure carries physical costs that conditional Landauer erasure does not—even though both protocols connect the same initial and final states. Any unitary $\hat{U}$ capable of achieving unconditional erasure must single-handedly transform *all* pre-erasure states to the same reset state $\hat{\rho}_{\text{reset}}^{\mathcal{S}}$, whereas each of the individual $\hat{U}_i$ applied in conditional erasure can be optimally selected to most efficiently transform a *single* pre-erasure state $\hat{\rho}_i^{\mathcal{S}}$ to $\hat{\rho}_{\text{reset}}^{\mathcal{S}}$.

3.3 On Protocol Classification

So far in this section, we proved general dissipation bounds for conditional and unconditional Landauer erasure that are transparently rooted in very fundamental considerations. Because only conditional erasure protocols utilize records (or copies) of pre-erasure system states, we again call the conditional erasure protocols of Sect. 3.1 ERASE WITH COPY protocols and the unconditional erasure protocols of Sect. 3.2 ERASE WITHOUT COPY protocols, respectively. The bounds we obtained for these two classes of protocols differ only in that the $k_B T \ln(2)$ Joule-per-bit "Landauer cost" appears in the dissipation bound for ERASE WITHOUT COPY protocols but does not appear in the corresponding bound for ERASE WITH COPY protocols. This supports both Landauer's limit for unconditional erasure and Bennett's claim that the Landauer cost can be avoided in erasure protocols that utilize copies.[4] These claims are widely accepted—the expectation of extreme energy efficiencies in reversible computing is, for example, largely predicated on the validity of Bennett's claim—but are not universally accepted.

In the remainder of this section, we address two challenges to the classification of erasure protocols as conditional or unconditional. The first challenge is to properly

[4]That Landauer himself regarded Bennett's claim as a "friendly amendment" to his erasure principle is evident in [34].

accommodate Bennett's distinction between "known data" and "random data" [11] within our approach, augmenting his definitions as appropriate and reconciling apparent contradictions between our respective conclusions regarding the associated dissipative costs of erasing random data. The second challenge is to clarify an apparent dependence of protocol classification—as conditional or unconditional—upon the specification of system boundaries.

3.3.1 Erasure of Known and Random Data Revisited

In responding to critics of his work [11], Bennett made a distinction between *known* and *random* data. He argued that Landauer erasure of known data is thermodynamically irreversible in general, but that it can be thermodynamically reversible—can in principle be achieved without paying the dissipative Landauer cost—if a copy of the data is available and is used to condition the protocol. He further argued that erasure of unknown data can be thermodynamically reversible without any conditioning of the erasure protocol on a copy of the data. Below we describe our own construal of the differences between physical states that bear known data, random data, and no data at all. We then compare and contrast our conclusions about the associated dissipative costs with those of Bennett, and show that an apparent contradiction between our respective conclusions about the dissipative cost of unconditionally erasing unknown and random data is resolved when account is taken of differences in our definitions.

With Bennett, we distinguish conditional ERASE WITH COPY from unconditional ERASE WITHOUT COPY operations, as discussed previously in Sect. 3.1. Also with Bennett, we distinguish physical states that encode known data from physical states that do not. In this work, however, we make an additional distinction that was not explicitly recognized by Bennett—the distinction between physical states that encode data and physical states that do not encode data at all. This third distinction matters. Without this distinction, there is nothing to differentiate a system that encodes data from a system that does not encode data *but that shares the same statistical state*. Differences in the respective options available to a would-be erasing agent[5] are thus obscured. With this distinction, however, two types of uncertainty are resolved for data-bearing systems. One is related to a would-be erasing agent's uncertainty in a system's state that originates from not knowing the data state in which the system has been prepared. The other is the uncertainty that remains when the (generally mixed) data state is known. These distinctions and their consequences are detailed below.

Specifically, we take a system to be encoding data if and only if it is prepared in *one* of the data states $\hat{\rho}_i^S$ *and* if there exists a record or copy of the data

[5]The "erasing agent", which we will also call the "operator", is the entity (organism or automaton) tasked with executing a state-reset protocols. This entity is the "knower" of a data-bearing system's preparation in the case of known data.

instantiated in a physical system that is external both to the system $\mathcal{S}$ and the observer-inaccessible environment $\mathcal{E}$. The requirement of an extant physical copy imbues the state of $\mathcal{S}$ with "aboutness"—building a physical answer to the question "data about what?" into the very definition of data—and more formally qualifies $\mathcal{S}$ as a bearer of physical information by the criteria explicated and defended in [35]. Simply "having a state" is for us not enough to encode data or bear information: that state must also be correlated with an external physical copy (or *referent*) that the system state encodes.

If a system bears data in the above sense, and the identity of the data state *is* known with certainty to the erasing agent, then we take the system to be encoding *known* data. If a system bears data and the identity of the data state is unknown to the erasing agent/automaton—or known only statistically with the statistics of the agent's knowledge reflected in the probabilities p_i (with $p_i < 1 \; \forall i$)—we take the system to be encoding *unknown* data. If, on the other hand, the system has *not* been prepared in any one of the data states, then we simply do not take the system to be encoding data; again, not all system states are data-bearing states. We now compare and contrast these notions with those used in Bennett's [11].

Consider first the conditional and unconditional erasure of known data. Our characterization of known data harmonizes with that of Bennett insofar as we both presume that the pre-erasure state of the system is one of the data states. That Bennett presumes as much is clear in the way he uses a Szilard engine[6]—a one-molecule gas in a cylinder is used to encode binary data (0 or 1) by confining the "left" or "right" half of the cylinder using a piston and partition—to distinguish the various cases [3, 11]. We differ with Bennett in that we would also require an external record of "left" or "right" before we would regard the engine to be bearing data in the first place. Yet, our conclusions regarding the dissipative costs of erasing known data with and without a copy—expressed in the bounds (3) and (12) of Sects. 3.1 and 3.2, respectively—align with those of Bennett as discussed below.

The conditional erasure bound (3) applies to erasure of known data *with* use of a copy: each term in the average (2) reflects the system-to-environment energy transfer resulting from resetting of *one of the* data states using a time-dependent potential specific to that state. The absence of a Landauer cost from (3) comports with Bennett's conclusion that erasure of known data can be reversible if a copy of the encoded data is available for conditioning of the reset protocol. In Bennett's Szilard engine examples, this corresponds to reversible isothermal expansion of the gas—necessarily preceded by insertion of a piston into what is known to be the "correct" side of the cylinder as per an external record of the data state—followed by unconditional piston insertion from the right and reversible compression to the left (reset) half of the cylinder.

[6]We will refer to Bennett's Szilard-engine examples solely to reveal similarities and differences in various notions used by Bennett and by us, and should not be taken to imply that any results of this paper depend in any way upon results obtained for Szilard engines, the applicability of classical thermodynamics to Szilard engines, or any assertion that microscopic realization of a Szilard engine could operate in the face of thermal fluctuations.

The unconditional erasure bound (12), on the other hand, applies to erasure of known data *without* use of a copy; each term in the average (2) reflects system-to-environment energy transfer resulting from resetting of one particular data state using a universal protocol that is used blindly to reset all data states. The presence of a Landauer cost in (12) comports with Bennett's conclusion that erasure of known data is irreversible without the existence *and use* of a copy to condition the protocol. In Bennett's Szilard engine examples, this corresponds to irreversible free expansion of the gas—achieved by unconditional removal of the partition without regard to the identity of the data state—followed by unconditional piston insertion from the right and reversible compression to the left (reset) half of the cylinder.

Consider next the definitions of unknown and random data and the erasure of such data, which is where our differences with Bennett arise. For us, a system bears unknown data if the pre-erasure state of the system is one of the encoding states but the identity of that state is known only statistically to the erasing agent: the preparation is random from the agent's point of view. An external record of the data exists, as we require for the system to bear data in the first place, so both ERASE WITH COPY and ERASE WITHOUT COPY protocols are possible for unknown data. For Bennett, however, a system bears random data if its statistical state coincides with the "average state" of the data states. No external copy is presumed, so there is no notion of conditioning the transformation of the system state to the reset state $\hat{\rho}_i^S$. The states that Bennett would take to be encoding random data are states that we would not take to be encoding data at all.

This difference between Bennett [11] and the present work is easy to see in the context of the Szilard engine. Suppose that the partition is absent, so the molecule wanders freely throughout the entire cylinder and would at any given time be equally likely to be found in the left and right halves. Bennett considers the gas to be encoding random data in such a scenario; he associates isothermal compression of the molecule in this state to the "reset side" of the cylinder—which can be done reversibly in principle—to be erasure of random data. We would, however, consider the molecule to be encoding *no data at all* in this scenario—known or unknown—even if the partition were to be inserted into the cylinder so the molecule is randomly trapped on one side or the other. Resetting of this molecule state could not, on our view, erase data or information, simply because the molecule does not encode data or bear information to begin with. For us, the molecule would not encode data unless it had been deliberately compressed into either the left or right half of the cylinder and trapped there by insertion of a partition *and* an external record of the "sidedness" had been retained. Resetting of such a state can be done reversibly—even in principle—only through operations that are conditioned on this record of the sidedness.[7]

[7]Conditioning on the sidedness does not require erasing agent knowledge of the sidedness, provided that the operation—applied "blindly" by the agent—acts on both the system and the copy in a manner that conditions the operation performed on the system.

Table 1 Erasure protocols and their reversibility for known, unknown, and random/no data

This work			Bennett		
Data	Protocol	Cost	Data	Protocol	Cost
Known	ERASE WITH COPY	Reversible	Known	ERASE WITH COPY	Reversible
	ERASE WITHOUT COPY	Irreversible	–	ERASE WITHOUT COPY	Irreversible
Unknown	ERASE WITH COPY	Reversible	–	–	–
	ERASE WITHOUT COPY	Irreversible	–	–	–
None	Reset	Reversible	Random	Erase	Reversible

Similarities and differences in Bennett's notions of known and random data and our own are summarized in Table 1, as are our respective conclusions regarding the reversibility of erasure. Our respective notions of known data are similar in spirit, and we agree that Landauer erasure of known data can in principle be reversible if a record of the data is used to condition the erasure protocol. Our notion of unknown data—with the identity of the system state (one of the data states) known only statistically to a would-be erasing agent, but registered in an external physical record—is not considered by Bennett. For such a scenario, our conclusions regarding the reversibility of erasure mirror those of known data: Landauer erasure can in principle be reversible for protocols in which actions taken on the system are conditioned on the external record of the encoded data, but are otherwise necessarily irreversible.[8] Finally, those cases where we do not regard the system to be encoding data at all—where there is no external physical record of the system state—are the cases that Bennett associates with the encoding of random data. Despite the difference in our "labeling" of this physical scenario—state-resetting of a non-data-bearing system for us and erasure of a random-data-bearing system for Bennett—we both conclude that the process can in principle be reversible for some situations. In all physical scenarios considered by us in this work and by Bennett in [11], we reach the same conclusions regarding dissipation but differ whether or how this dissipation is associated with data/information erasure.

We conclude this discussion with a few final remarks on distinction between scenarios that Bennett would take to initially encode random data and that we would take to encode no data at all. We specifically remark on two distinct physical scenarios—germane to the Szilard engine—where we would maintain this distinction even though the statistical state of the gas molecule is identical in the two scenarios. One scenario has the molecule compressed into the left or right side and held there by the partition, but with the "sidedness" completely unknown to the erasing agent, and the other is with the partition absent so the molecule wanders freely throughout the chamber. The first scenario corresponds to encoding of unknown data by our definition, and the second to no encoding of data at all by our definition and of random data by Bennett's.

[8]Note that reversible erasure of unknown data with a copy requires operations, applied blindly by the erasing agent, that access the copy without destroying it. See Sect. 4 and [36].

The comparison begs the question: if all physical properties of a system depend on the system state, how can there be any meaningful physical distinction between two scenarios involving identical systems prepared in identical statistical states. How can one be data-bearing and the other non-data-bearing? Absent an answer to this question, it would indeed seem that the distinction is meaningless. This would imply that all systems bear data, that no systems bear data, or that the designation of a system as data-bearing or non-data-bearing is arbitrary.

We reject all three possibilities, because we do not regard the molecule's status as a data-bearer as being dependent on the state of the molecule alone: it depends on the *joint* state of the molecule and some non-environmental referent system about which the molecule encodes data. If the molecule state is correlated to the referent state—in other words if the referent holds a *record* or *copy* of the molecule state—then the molecule state holds data about that referent (and vice versa). If there is no such external referent that holds a copy of the molecule state, then the molecule does not bear data about that referent. The local molecule states can be identical in scenarios where the molecule and referent are and are not correlated, i.e. the joint referent-molecule states can differ crucially when the local molecule states are identical. This resolves the difficulty, clearly distinguishing on physical grounds those physical states that encode data and bear information from those that do not.

To illustrate the resulting disambiguation in a familiar context, consider the combination of a Szilard engine and one-bit memory that Bennett used to exorcise Maxwell's Demon in [3]. If the molecule state is correlated to the state of the memory with the partition inserted (Fig. 12(c) in [3]), then the molecule is in a data bearing state. If the molecule state is not correlated to the state of the memory, either with the partition inserted (Fig. 12(b) in [3]) or removed (Fig. 12(e) in [3]), then the molecule is not in a data-bearing state. The statistical state of the molecule alone is identical for all three of these cases, and the memory (referent) states are also identical in the first and third ((c) and (e)), but the corresponding joint states differ crucially in that the molecule and memory are correlated—the molecule encodes data about the memory and vice versa—only in the first case (c). The problematic ambiguity has thus been removed.

3.3.2 System Boundaries and Protocol Classification

We now discuss an ambiguity in classification of erasure protocols as either conditional or unconditional. Specifically, we will show that choice of the boundaries that define "the system" in statements of the Landauer and Landauer-Bennett limits can affect this classification—and thus the interpretation of these limits—illustrating why great care must be taken in claiming "violation" of the Landauer limit.

Consider the following scenario involving a composite system $\mathcal{RS}$. Suppose that the subsystem $\mathcal{S}$ interacts with the subsystem $\mathcal{R}$, which holds a durable copy of the initial state of $\mathcal{S}$ *before, during, and after* resetting of the state of $\mathcal{S}$. Suppose also that the forces *applied* to reset $\mathcal{S}$ *are not* conditioned on the initial state of $\mathcal{S}$. A familiar example of this nature, used below to illustrate the classification ambiguity for such a scenario, is the ERASE WITH COPY operation via Bennett clocking in the quantum-dot-cellular automata (QCA) [13].

The primitive element in six-dot QCA is a three-state cell that uses two polarized cell states (here $\hat{\rho}_0$ and $\hat{\rho}_1$) to encode a bit value and a third "null" state (here $\hat{\rho}_{\text{reset}}$) as a reset state. State transitions required to implement communication and logic in a QCA cell array are enabled by intercell coupling and by synchronized clocking fields unconditionally applied to selected regions of the array. Lent and co-workers have identified the clocked ERASE WITH COPY operation as the enabling operation for reversible computation in the QCA nanocomputing paradigm [13].

We specifically consider an idealized erasure scenario involving two adjacent QCA cells—a "memory" cell $\mathcal{S}$ and a "copy" (referent) cell $\mathcal{R}$—that are placed within close proximity to one another so they interact via a local coupling field. Initially, at $t = 0$, the same bit value is encoded in both $\mathcal{S}$ and $\mathcal{R}$. During the interval $0 \leq t \leq t_{\text{er}}$, a standard time-dependent clocking potential $V_{\text{clk}}(t)$— selectively but unconditionally applied to $\mathcal{S}$—transforms $\mathcal{S}$ from its initial state ($\hat{\rho}_0^{\mathcal{S}}$ or $\hat{\rho}_1^{\mathcal{S}}$) at $t = 0$ to its reset state $\hat{\rho}_{\text{reset}}^{\mathcal{S}}$ at $t = t_{\text{er}}$ without affecting the copy $\mathcal{R}$. The two possible state transitions of the joint system for this scenario are

$$\hat{\rho}_0^{\mathcal{R}} \otimes \hat{\rho}_0^{\mathcal{S}} \rightarrow \hat{\rho}_0^{\mathcal{R}} \otimes \hat{\rho}_{\text{reset}}^{\mathcal{S}}$$

and

$$\hat{\rho}_1^{\mathcal{R}} \otimes \hat{\rho}_1^{\mathcal{S}} \rightarrow \hat{\rho}_1^{\mathcal{R}} \otimes \hat{\rho}_{\text{reset}}^{\mathcal{S}}.$$

This resetting of $\mathcal{S}$ may at first wash seem to be a conditional ERASE WITH COPY protocol, since $\mathcal{S}$ is reset *while interacting with* the cell $\mathcal{R}$ that holds a copy of $\mathcal{S}'s$ initial state. Conditioning of the forces transforming $\mathcal{S}$ on the state of $\mathcal{R}$, and thus on its own initial state, is "built in" to the joint system $\mathcal{RS}$ throughout the erasure operation. But the state reset is achieved by applications a standard time-dependent clocking potential, *blindly* applied to $\mathcal{S}$ with no regard for its initial state. This might instead suggest classification of this operation as an unconditional erasure protocol. So which is it? The answer matters, since conditional and unconditional erasure protocols obviously carry different physical costs and different physical costs should not result for identical processes on identical systems just because external observers classify them differently. This is explained below.

We argue below that one could appropriately classify the operation as either conditional or unconditional from the erasing agent's point of view, depending on whether the memory subsystem $\mathcal{S}$ or the memory-copy composite $\mathcal{RS}$ is taken to play the role of "the system" referred to in statements of the Landauer and Landauer-Bennett limits—the system from which data or information is erased. Yet, we further argue, only identification of the system as $\mathcal{S}$ alone admits identification of the protocol as Landauer erasure—i.e. to a resetting of the state of "the system."

Suppose first that we take subsystem $\mathcal{S}$ to be "the system." The ever-present influence of $\mathcal{R}$ shows up as a *static* contribution $V_i^{\mathcal{S}}|_{\text{copy}}$ to the time-dependent external potential $V_i^{\mathcal{S}}(t)$. The static force exerted on $\mathcal{S}$ by $\mathcal{R}$ is determined by the state of $\mathcal{R}$, which depends in turn on the initial state of $\mathcal{S}$, so the static contribution

$V_i^{\mathcal{S}}|_{\text{copy}}$ to $V_i^{\mathcal{S}}(t)$ is the only component of the full Hamiltonian that depends on the initial state of $\mathcal{S}$. The remainder of $V_i^{\mathcal{S}}(t)$ is the blindly applied, time-dependent external clocking potential $V_{\text{clk}}^{\mathcal{S}}(t)$. The appropriate Hamiltonian is thus

$$\hat{H}_i^{\mathcal{SE}}(t) = (\hat{H}_{\text{self}}^{\mathcal{S}} + V_i^{\mathcal{S}}|_{\text{copy}}) + \hat{H}_{\text{self}}^{\mathcal{E}} + \hat{H}_{\text{int}}^{\mathcal{SE}} + V_{\text{clk}}^{\mathcal{S}}(t)$$

for $0 \leq t \leq t_{\text{er}}$. Because the corresponding $\hat{U}_i$ depend on the initial state of $\mathcal{S}$ (through the $V_i^{\mathcal{S}}|_{\text{copy}}$), and because the (automatically) conditional application of these $\hat{U}_i$ resets the state of "the system" $\mathcal{S}$, the QCA ERASE WITH COPY operation qualifies as conditional Landauer erasure of (generally unknown data) from $\mathcal{S}$. With minor modifications of the proof from Sect. 3.1, properly accounting for dependence of $\mathcal{S}'s$ "effective" self-Hamiltonian $\hat{H}_{\text{self}}^{\mathcal{S}} + V_i^{\mathcal{S}}|_{\text{copy}}$ on its initial state, the bound (3)—our generalization of the Landauer-Bennett limit—is recovered unscathed. No $k_B T \ln(2) \Delta I_{\text{er}}$ term appears in the lower bound on the erasure cost, consistent with the expectation of Ref. [13].

Next, suppose next that we instead take "the system" to be $\mathcal{RS}$. Interactions between $\mathcal{S}$ and $\mathcal{R}$ are built in to "the system" and its *self*-Hamiltonian $\hat{H}_{\text{self}}^{\mathcal{RS}}$. The total Hamiltonian is

$$\hat{H}_i^{\mathcal{RSE}}(t) = \hat{H}_{\text{self}}^{\mathcal{RS}} + \hat{H}_{\text{self}}^{\mathcal{E}} + \hat{H}_{\text{int}}^{\mathcal{RSE}} + V^{\mathcal{RS}}(t)$$

for $0 \leq t \leq t_{\text{er}}$. The external "forcing" term $V^{\mathcal{RS}}(t)$ now reflects only the blindly applied clocking potential, which is independent of the initial state of $\mathcal{RS}$ so the corresponding reset protocol is unconditional. The protocol *is not*, however, a *Landauer erasure* protocol, since only the state of the subsystem $\mathcal{S}$ is reset. The final state of $\mathcal{RS}$—"the system" here—is *not* reset since the initial state of subsystem $\mathcal{R}$ remains intact throughout the operation. The bound (12) obtained for unconditional *Landauer* erasure therefore does not apply to this unconditional "subsystem reset" operation. The appropriate lower bound for *this* operation, obtained as was (3) but with the $\Delta\langle E_i^{\mathcal{E}}\rangle \geq -k_B T \ln(2) \Delta S_i^{\mathcal{RS}}$ evaluated for the state transitions of the composite $\mathcal{RS}$, also returns the form (3) (since $\Delta S_i^{\mathcal{RS}} = \Delta S_i^{\mathcal{S}}$ here). This is as it should be, since the two scenarios considered here are physically identical and are only interpreted differently.

Thus, an ERASE WITH COPY protocol characterized by a "built-in" system-copy interaction (as in the six-dot QCA example above) is indeed properly classified as a conditional Landauer erasure protocol, even if it seems "unconditional" in the sense that it is driven by the unconditional application of a standard external clocking potential. The potentials $V_i^{\mathcal{S}}(t)$ in this erasure protocol include both an external static "conditioning" bias—resulting from persistent interaction of the system with the external nonvolatile copy of its initial state—and a time-dependent contribution that enables the state transformation but is *not* conditioned on the initial state. The external static bias conditions the resetting of the state of subsystem $\mathcal{S}$—regarded as "the system" in this case—and there is no minimum energy cost related specifically to erasure of information as expected: the result is consistent with the Landauer-Bennett limit, which would apply to this case.

However, with the boundary of "the system" simply redefined to include subsystems $\mathcal{R}$ *and* $\mathcal{S}$, the external potential includes only the unconditional component. Unconditional application of this potential resets the state of $\mathcal{S}$ but there no Landauer cost appearing in the corresponding dissipation bound, which may seem to suggest that the Landauer limit has been violated. There is, however, no violation, as Landauer erasure of "the system" has not achieved with the "system" defined as $\mathcal{RS}$, since the state of subsystem $\mathcal{R}$ remains intact. A mistaken interpretation that the Landauer limit has been violated would follow from an inconsistency in defining "the system"—as $\mathcal{S}$ for classification of the protocol as Landauer erasure (since it is $\mathcal{S}$ that is reset) and as $\mathcal{RS}$ for classification of the protocol as unconditional (since the externally applied potential is not conditioned on the initial state of $\mathcal{RS}$).[9]

4 Discussion: On Theoretical Methodology

In Sect. 3, we obtained bounds for the energy cost of Landauer erasure in physical systems under various assumptions about the presence and use of external copies of the to-be-erased data. The bounds (12) and (3), which generalize the Landauer and Landauer-Bennett limits respectively, apply to erasure protocols that always begin with the data-bearing system prepared in a state that belongs to a fixed set of N orthogonal and generally mixed "encoding" states, and always end with the system in a standard, generally mixed reset state. Our aim has been to deliberately prove these inequalities in a manner that immunizes them from methodological objections that have been leveled against other proofs, arguments, and demonstrations of the Landauer limit. In this section, we reiterate some of the most prominent objections and show how they are sidestepped in our approach.

For present purposes, we group most of these objections in three categories. The first category concerns proper application of thermodynamics to the erasure problem, which is significant because proofs and illustrations of the Landauer and Landauer-Bennett limits have relied most heavily on thermodynamic thought experiments involving classical gasses in cylinders and protocols that employ

[9]One would be similarly mistaken to see a violation of the Landauer limit in the nanomechanical OR gate recently reported by Lopez-Suarez, Neri, and Gammaitoni [37]. They report that, in their experiment, two distinct configurations of electrode charges (corresponding to two different binary input combinations "01" and "10") similarly displace the position state of a nearby nanopillar tip (same logical output) with energy dissipation less than that Landauer limit. If this does indeed correspond to implementation of a logically irreversible (sub-)function, as the authors claim, there is no violation of the Landauer limit. Regarding the nanopillar as the "the system" in this scenario, not including the electrode tips that must remain charged to hold the nanopillar tip in its "merged position", the "state merging" protocol is conditional and the Landauer-Bennett limit (*not* the Landauer limit) would apply. No Landauer cost would be expected. That a different nanopillar tip position results for electrode charges corresponding to input "11"—which should yield the same physical output as inputs "01" and "10" in a faithful physical implementation of an OR gate—is a separate worry.

frictionless pistons, ideal removable partitions, and familiar reversible and irreversible thermodynamic processes (e.g., free expansion, isothermal compression and expansion). The second category of objections concerns the role—even the necessity—of copies and conditioning in avoiding the Landauer cost in erasure. The third category concerns the idealized nature of the processes typically used in proofs of the Landauer and Landauer-Bennett limits, the "indirectness" of these proofs in some cases, and the implications of idealization and indirectness for the validity and interpretation of bounds so proven.

Objections of all three sorts figure prominently in John D. Norton's 2011 paper "Waiting for Landauer" [16], which critiques thermodynamic proofs Landauer's Principle. Right out of the gate, Norton summarizes his primary concerns:

> Landauer's Principle asserts that there is an unavoidable cost in thermodynamic entropy creation when data is erased. It is usually derived from incorrect assumptions, most notably, that erasure must compress the phase space of a memory device or that thermodynamic entropy arises from the probabilistic uncertainty of random data. [16]

He then acknowledges that "recent work seeks to prove Landauer's Principle without these assumptions" but promises to show that they fail because the assumed processes can be combined in ways that would obviously violate the Second Law and "worse" that these processes neglect thermal fluctuations. He also promises to show how "concrete proposals" for reversible expansion of single-molecule gasses are "fatally disrupted" by thermal fluctuations that "can only be overcome by introducing entropy creating, dissipative processes." His overall conclusion is that "we still await a cogent justification of Landauer's Principle and that present efforts to demonstrate it are proceeding in an incoherent framework" [16].

As perhaps the most forceful and clearly articulated critique of thermodynamic treatments of Landauer's Principle, Norton's paper provides a good vehicle for framing objections to such proofs. Below we draw heavily from Norton's paper [16] and from the ensuing exchange of papers [17–19] with Ladyman and Robertson to introduce these objections and some responses. Our objective is not to mount either a point-by-point rejection or defense of thermodynamic proofs; we endorse Norton on some points and differ with him on others. Rather, it is to show how objections to thermodynamic proofs of the Landauer and Landauer-Bennett limits are sidestepped in our quantum-dynamical proofs of Sect. 3.

4.1 Application of Thermodynamics to Erasure

We start with four related objections to thermodynamic treatments of erasure cost, all emphasized by Norton and all couched in terms of the Szilard engine.[10]

[10] While Norton has separate objections to the Szilard engine as an idealization, he accepts that "it is taken to capture the essential thermodynamic features of a more realistic one-bit memory device in a heat bath" [16] and states his thermodynamic objections in this spirit.

The first is the objection to the way phase-volume arguments have been used in thermodynamic treatments of erasure. When a partition is inserted into a Szilard engine, with the molecule trapped in either the right or left half of the chamber with equal probability, thermodynamic proofs of the Landauer limit take the molecule's accessible phase volume to be the phase volume associated with the full chamber (since it may occupy either side). Norton objects that the phase volume is actually halved when the partition is inserted, undermining thermodynamic proofs of the Landauer limit that would locate the erasure cost in doubling of the environment entropy that results from halving of the molecule's phase volume during resetting:

> The error of the proof is that the molecule, prior to erasure, is not associated with an accessible phase volume that spans the entire chamber. It will assuredly be in one half only. Which half will vary from occasion to occasion, but it will always be one half. As a result, the erasure operation does not need to reduce accessible phase volume at all; it merely needs to relocate the part of the phase space accessible to the molecule. [16]

Such objections are sidestepped in the quantum dynamical proofs of this work, which do not invoke classical phase-volume arguments or analogous arguments that would render the system's state space time-dependent or otherwise ambiguous.[11] The pre-erasure states of a data-bearing system—our analog to the "left" and "right" trapped molecule states—have support on orthogonal subspaces of the full-system Hilbert space, as required for unambiguous encoding of data, but nothing in the Hamiltonian limits access to the full system state space before, during, or after erasure.

Second, Norton objects that in proofs making use of ensembles "the 'random' data state is treated as if it was the same as a thermalized data state" but that "the random data state and the thermalized data state are not thermodynamically equivalent." We agree with Norton on this point, and do not presume any such equivalence in this work. Indeed, we have explicitly distinguished scenarios where a system encodes random data from those where a system encodes no data at all (*cf.* Sect. 3.3), the latter of which corresponds to Norton's "thermalized data state." While we further agree with Norton that the fact that "(w)e do not know which half of the phase space is accessible in the case of random data is irrelevant to the device's thermodynamic properties," it does not follow that such knowledge

[11] Phase-volume arguments are themselves artifacts of the idealization of an absolutely impenetrable partition. An arbitrarily small pinhole in the partition—a pinhole so small that only one molecule could fit through it and would almost never do so—is enough to give the molecule access to the full phase volume even when it is (temporarily) trapped on one side or the other by the partition. The presence of such a pinhole would be sufficient to undermine phase volume arguments even when the operation over finite time scales is unaffected, and any quasi-static treatment that would require the molecule to always be in equilibrium with its surroundings. With the pinhole present, a particle trapped on one side of the partition would end up on either side with equal probability when it has fully equilibrated. The same is true for a quantum particle in a symmetric double potential well with a high *but finite* potential barrier in the center and interacting with a heat bath. Any particle state initially localized in one well or the other—necessarily a non-equilibrium state—will ultimately equilibrate to a thermal state that does not favor occupation of either side of the chamber by the molecule.

is irrelevant to the minimum dissipative price that is paid when agents implement reset protocols with and without the benefit of this knowledge. Here we side with Ladyman and Robertson:

> Norton questions why we should think that whether we know about which state a device is in affect matters. But it does matter: thermodynamics is about the properties of matter and how we can exploit or use these properties to do work. Knowing what state a device is in changes which operations you are able to perform. [17]

Indeed, we showed explicitly in Sects. 3.1 and 3.2 that the Landauer cost arises only when the agent implementing an erasure-by-reset protocol does not know (or have access to a record of) the pre-erasure data state of a device and make good use of this knowledge or record. This agent simply does not have as good or as many options for implementing the protocol as does an agent who does possess this knowledge (or who has access to such a record).

Third, Norton argues that logical specification of the erasure function should not involve probabilities. He asserts that the "introduction of probability is routinely assumed benign in physical analysis when some variable has an indeterminate value" but that

> It is not benign since it adds non-trivial structure to the indeterminateness of a variable and can induce egregious inductive fallacies, as shown in Norton (2010). The real seat of the (Shannon entropy formula) is this probability distribution and, absent cogent justification of the introduction of the probability distribution, the entropy change associated with erasure by (the Shannon entropy formula) is merely an artifact of a misdescription of the indeterminateness of data. [16]

The encoding probabilities p_i that are used in this work, and that appear in the Shannon information formula that emerges in the dissipation bounds for unconditional erasure protocols, do not represent the "indeterminateness of data" at the root of Norton's concern. Here they enter straightforwardly and unproblematically as weightings in a garden-variety average of costs associated with a set of individual *deterministic* physical processes connecting *specified* initial and final states. Specifically, these averages are environmental energy changes incurred in transformations of various initial pre-erasure system states to final reset states—both specified—for individual systems in large ensembles or individual erasure trials on a single system. Such an average is unavoidable if we are to associate a single number—an "energy of erasure" as appears on the left-hand side of Landauer's inequality—with protocols that, like erasure, involve "families" of physical processes[12]—each of which links a single initial and final state. In erasure, there is one such deterministic process for each initial encoding state, there are generally different dissipative costs associated with resetting of each of these states, and the relative frequencies of systems in the ensemble that are prepared in the various initial states (and that thus result in execution of the various processes) generally differ. Hence the necessity for averaging, even though no states or processes are indeterminate at the level of individual systems or trials. The weighting probabilities that enter the proof through

[12]This terminology is due to Ladyman [38].

this straightforward averaging show up later in the Shannon entropy formula that emerges mathematically in the absence of conditioning. Specifically, in obtaining the unconditional erasure bound (12), the statistical state $\hat{\rho}^{\mathcal{S}} = \sum_i p_i \hat{\rho}_i^{\mathcal{S}}$ emerged when the conditional erasure bound (3) was specialized for the case of unconditional erasure and the sum over the $\hat{\rho}_i^{\mathcal{S}}$ was "absorbed" into time evolution transformation. The objection is thus sidestepped.

Fourth and finally, Norton notes that some thermodynamic proofs that focus on a system's information theoretic entropy and its reduction in erasure is assumed to correspond to a reduction of the device's thermodynamic entropy [16]. This continues a longstanding debate on the connection between information and entropy [39], perhaps most recently extended by [40]. No such equivalence has been presumed in this work: the von Neumann entropies of the individual pre- and post-erasure states and the properties of von Neumann entropy are used as vehicles to obtaining the energy bounds, but they are at no point assumed to have information-theoretic significance or to be equivalent to thermodynamic entropies. The Shannon entropy appearing in the bound unconditional erasure bound (12) again emerged mathematically in a proof that started with a straightforward averaging over definite processes, and was not put in "by hand" or associated with any "illicit" ensemble construction. Absolutely no a priori assumption was made that pre-erasure states $\hat{\rho}_i^{\mathcal{S}}$ of the collection of systems can be legitimately combined into a density operator $\hat{\rho}^{\mathcal{S}} = \sum_i p_i \hat{\rho}_i^{\mathcal{S}}$ that can then be used to represent an ensemble in the thermodynamic sense. Such an assumption had indeed been made in the literature, and has been challenged (e.g., [4, 12]), but is not made here. This objection is sidestepped as well.

Many other objections to thermodynamic proofs and applications Landauer's limit are possible, and some have been stated by Norton and others. Some stem from use of equilibrium expressions for free energy to derive Landauer's limit, and others from the use of this limit—when regarded as a direct result of the Second Law—to justify results about adherence to the Second Law by Maxwell's Demon. Such objections are sidestepped here as well. We have avoided use of any "off the shelf" results from thermodynamics—phase-space conservation, equilibrium free energy expressions, the Second Law—in proving the Landauer and Landauer-Bennett limits, and have made no assumptions regarding the equivalence of thermodynamic, information-theoretic, and von Neumann entropies. The bounds (3) and (12) derive exclusively from the properties of global Schrodinger evolution, established properties of von Neumann entropy, distinguishability of the various states used to encode data, and the assumption that, *prior* to all state resets, *the environment* is in the thermal state $\hat{\rho}_{\text{th}}^{\mathcal{E}} = Z^{-1} \exp[-\hat{H}^{\mathcal{E}}/k_B T]$ (where Z is the partition function). The data-bearing system itself is at no point assumed or required to be in equilibrium with its environment or in any state with a well-defined temperature.

4.2 The Roles of Conditioning and Copies

Next, we take up objections related to the roles of conditioning and use of records of the to-be-erased data to lower the dissipative cost of erasure.

Norton [16] acknowledges a distinction between unconditional and conditional erasure protocols, if not by those names, and the possibility of lowering erasure cost through conditioning. This is evident is his discussion of "erasure by thermalization." By this he means irreversible free expansion by partition removal followed by reversible isothermal compression to the reset state. He calls the initial step of removing the partition "ill-advised," and correctly points out that the irreversibility of this particular protocol "does not show that all possible erasure processes must create thermodynamic entropy." He further notes that defenders of the Landauer limit require "removal of the partition, or something like it, for the process cannot 'know' which side holds the molecule on pain of requiring further erasure of that knowledge" [16].

We completely agree that partition removal "or something like it"—something *unconditional*—is required when reset is achieved without knowledge or other records of the pre-erasure state, i.e. those very scenarios to which the Landauer limit applies. This is not inconsistent, however, with the recognition that such processes can be avoided by "relocating" the "part of the phase space accessible to the molecule." This relocation must be achieved by operations that are conditioned on knowledge or durable records, and belong to those scenarios to which Landauer-Bennett limit applies.

Surprisingly, Norton seems to deny the necessity of records for achieving reversible erasure. In a Maxwell's Demon cycle that does not require the usual "operator" knowledge of which side the molecule is on (or external records of the same), he invokes what he calls a *Dissipationless erasure* operation—a would-be reversible operation that requires no operator knowledge of the pre-erasure molecule state or any other durable, pre-existing copy of the same. Norton argues that this operation is reversible since it is composed from *Detect and trigger* and *Shift* operations, both of which he takes to be reversible.

We (and certainly others) would argue, however, that this operation would require creation of an external copy of the system's pre-erasure state that must be left behind. This matters because, in a cycle, this copy would have to erased in a later step using an unconditional—and thus necessarily dissipative—erasure process. Norton objects:

> The claim is unsustainable. The triggered process can proceed without the continued existence the triggering data; all that it needs is for the data to exist at the time of the triggering and the presumption that the processes can proceed independently once triggered. If it helps, imagine that the process triggered is carried out by a physically distinct robotic machine. The device's sole function is to perform this one process without needing any further data input; it operates autonomously once triggered; and it is programmed to return itself to its unique ready state as its last step. [16].

Norton does not, however, say why the robotic device's return to "its unique ready state in its last step" is does not amount to an unconditional—and thus dissipative—Landauer erasure operation. As Ladyman and Robertson put it:

> Norton claims that the idea of 'controlled operation' from a degree of freedom to itself is vindicated since a robot may be posited to enact it. However, contrary what Norton supposes, any such robot would have internal degrees of freedom that would store the state of the control bit (the state of the device), thereby effectively remembering the state of the target bit, and hence requiring resetting for a cycle to be completed. It is also arguable that such 'controlled operations' of a degree of freedom on itself are not viable because any operation would require an auxiliary system whose internal state would determine which operation was performed. For example, a piece of paper that says 'destroy me' cannot read and destroy itself but would have to be destroyed by a system that read it, that is copied it, first. [17]

We side with Ladyman and Robertson here as well. More importantly for present purposes, however, we note that our proofs of the Landauer and Landauer-Bennett limits are not tied to arguments involving Szilard engines or Maxwell's Demons. They are very general, distinguishing at a realization-independent level between protocols utilizing reset operations that are conditioned on pre-erasure system states (the $\hat{U}_i$ of Sect. 3.1) and reset operations that are not (the $\hat{U}$ of Sect. 3.2). Nor do our arguments rely upon anthropomorphism of the "operator," which Norton considers grounds for dismissal of arguments that require it:

> Th(e) mistaken view persists, as far as I can see, because it is easy to anthropomorphize the erasure device as a little man who must always record what he is doing and then erase his records at the end. Absent that anthropomorphism, it is hard to see how the mistake can be sustained.

In this work, the absence or presence of conditioning is reflected in the structure of impersonal Hamiltonian operators that act on the systems in questions and that provably have dissipative consequences that depend on conditioning. The notions of knowledge, records, and conditioning have been physically cast, and need not be anthropomorphized. The objection is sidestepped.

4.3 Indirectness and Interpretation

We next consider objections to what Norton has called the "indirectness" of some proofs of the Landauer and Landauer-Bennett limits, and to the interpretation of arguments against their achievability as arguments against their validity.

Norton points out that concrete demonstrations of the Landauer limit—what he dubs "direct approaches" to their proof—may be based on inefficient erasure protocols like "erasure by thermalization" in a Szilard engine. He argues that such a demonstration, based as it is on a particular realization and protocol, cannot serve as a proof for a general principle: the principle would then "(depend) essentially on a poor choice of a convenient, but dissipative erasure procedure and (would) not derive from some essential feature of erasure itself" [16].

We would agree, and see in this observation a motivation to seek more general indirect approaches that, like the present work, aim to identify and theoretically capture the "essential feature of erasure itself" and quantify its lawful physical consequences. Norton, however, expresses skepticism about any such possibility.

For example, in criticizing the "indirect" approach of Ladyman, Presnell, Short, and Groisman [41] (hereafter LPSG), and its positing of what he calls a "statistical form" of the Second Law, he seems to take issue with indirect approaches in general:

> The attraction of [LPSG's] method is that we can leave the details of the erasure process undefined and seek a result that will apply to all erasure procedures. The weakness is that it automatically precludes illumination of the origin of the entropy cost of erasure; we can only infer that it must be there if the suppositions obtain. [16]

He argues that the suppositions do not in fact obtain—which would obviously be problematic for LPSG—but he continues:

> In positing a statistical form of the second law of thermodynamics at the outset, LPSG ...make no attempt to ground the law in the underlying physical properties of the systems to be investigated. [16]

The question of whether LPSG have posited an invalid physical law, as crucial as it is to the validity of any argument they make that depends upon it, is separate from the more general question of whether indirect approaches are tenable. Although Norton discusses the two questions together, he is clearly faulting the indirectness of the approach for neglecting the "underlying physical properties" of the systems question in favor of lawful relations that hold independent of all but the most essential properties.

In our view, however, if the particulars of various realizations *can be* subsumed under more general relations—immediate consequences of physical law that successfully isolate those physical properties "essential to erasure itself"—then all the better for the pursuit of results like the Landauer and Landauer-Bennett limits. These inequalities, interpreted as inviolable fundamental limits, quantify the unavoidable costs essential to erasure and separates them from all subsidiary costs associated with implementation of the required processes in particular systems. There is nothing unfamiliar about the pursuit of such generality in other physical contexts; it is what gives physical laws and their immediate consequences such descriptive power.

Norton's distaste for indirect proofs seems to be that they cannot establish the *achievability* of the Landauer limit. In his discussion of the literature, for example, he criticizes the neglect of "further thermodynamic costs" (beyond information erasure) "that might compromise the core idea to be protected: that ineliminable dissipation only arises through processes that physically implement logically irreversible functions." It is clear from this passage—from the use of "only"—that Norton is criticizing Landauer's limit as a claim of achievability, i.e. a claim that there are no ineliminable dissipation sources beyond logical irreversibility.

There are, of course, additional sources of dissipation in any realization. In some of the literature critical of Landauer's limit, however, such an acknowledgement metastasizes into a claim that the Landauer or Landauer-Bennett limit is *incorrect*

or *invalid*. While Norton himself stops short of this in claiming that "we still await a cogent justification for Landauer's Principle" [16], others (e.g., [25]) do not.[13] But validity and achievability not the same thing—interpretation of a bound as a limit whose violation would contradict physical law is distinctly different from its interpretation as achievable limit.

The root of the disconnect is a conflation of expectations for what can be obtained from high-level fundamental physical descriptions on the one hand and from explicit physical models of particular realizations on the other. High-level physical descriptions—like those used to describe erasure protocols in this work—apply to broad classes of protocols executed on any physical system that support the associated states and state transformations. They can be used to obtain fundamental bounds that are inviolable because they express direct consequences of bedrock physical law. They may or may not be achievable in any given realization, with the possibility of achievability or near-achievability depending on the particulars of that realization, but their inviolability is realization-independent.

By contrast, explicit physical models—like the device and circuit models that dominate engineering descriptions of computing components and circuits—are low-level descriptions of specific implementations of protocols executed on particular physical device realizations. Such descriptions can be and are used to obtain numerical estimates for highly specific situations. Their accuracy is determined by the extent to which all relevant phenomena have been captured in the model and all mathematical equations have been solved correctly and to sufficiently high resolution. The associated demands account for the complexity of such descriptions.

Fundamental physical descriptions are thus too general to provide accurate numerical estimates of resource requirements for particular realizations of devices operated in specified ways under particular conditions. Explicit physical device models of particular realizations are too specific to reveal fundamental limits that apply to all realizations. There is nothing contradictory about this, provided that the two are not conflated with one another. Nor is there anything contradictory about "simple, sharp, principled expression(s)[14]" for fundamental limits on quantities that could be accurately estimated in particular situations only through highly complex calculations.

No mechanical engineer would, for example, expect an accurate numerical estimate of a real heat engine's efficiency from a simple two-variable model. She would likely rely on a complex mathematical model of the engine, specialized to the engine design of interest by numerous parameters and solved via computer simulation. However, that same engineer would argue strenuously against the possibility that a real heat engine of that particular design—*or of any other design*—

[13] For example, in discussing an inequality presented as an inviolable fundamental bound with no claim of achievability, the authors of [25] state that "It should be emphasized that the greater-than-or-equal-to sign—rather than a greater than sign—is very important because the equality must represent a physical possibility, at least at the conceptual level."

[14] Norton questions whether it is possible to express limits to computation in a "simple, sharp, principled expression(s) [16].

could ever operate at an efficiency exceeding the value she would obtain from simple substitution of two temperature values into Carnot's formula for the limiting efficiency of a heat engine. In doing so, she would simply be recognizing the differing nature of the efficiency estimate obtained from her computer simulation of a particular heat engine and the Carnot limit on the efficiency of any heat engine.

This engineer would, however, raise eyebrows if she were to claim that the Carnot efficiency limit is invalid because she doesn't expect real heat engines to ever achieve this limit or because it does not provide accurate numerical efficiency estimates for particular engine realizations. Yet, much of the criticism of the Landauer and Landauer-Bennett limits is of this nature: arguments against the achievability of these limits are presented as arguments against their validity, conflating achievability with inviolability (e.g., [25]). Perhaps this is, at least in part, a historical result of the fact that Bennett's early papers on what we have here called the Landauer-Bennett limit also introduced the notion of reversible computing. The highest technological hopes for reversible computing *do* presume at least near-achievability of this limit, not just its inviolability. Whatever the case, it seems that debates about these limits would be much less confusing if validity as inviolable bounds and achievability were clearly distinguished and debated separately.

We have tried to clearly maintain this distinction in the present work. We have provided indirect proofs of the Landauer and Landauer-Bennett limits, and argued for their validity as inviolable fundamental limits. We have made no claim for the achievability of these limits, simply because our proofs cannot establish as much. The $k_B T \ln(2)$ per bit Landauer cost appearing in the Landauer limit represents a lower bound on *the contribution* of logical irreversibility to the dissipative cost of erasure—a dissipative contribution that *can* be expressed as a "simple, sharp, principled expression"—without denying that additional factors (e.g., parasitic losses, thermal fluctuations) might threaten the achievability of this limit in principle or practice and may be complicated or impossible to estimate with high accuracy. The "essential feature of erasure" responsible for the Landauer cost is simply the loss of distinguishability of data-bearing states upon resetting. The essential dissipative consequences of this loss of distinguishability are transparent consequences of physical law, and are independent of the "underlying physical properties" of specific realizations of data-bearing systems. States of these systems are represented by density operators, which do not require identification of the state variables used to differentiate the encoding states (e.g., position) but capture the distinguishability of these states. Dynamical processes that reset system states are governed by lawful unitary evolution of the system and its surroundings include the structure but not the details of these interactions. The unitary operations $\hat{U}_i$ and $\hat{U}$ used in our proofs represent *all* conditional and unconditional unitary operations that can perform the tasks that their respective protocols require them, i.e. that generate the required initial-to-final state resetting transformations.

5 Consistency with Experiment

Finally, we consider the consistency of the conditional and unconditional erasure bounds proven here—and thus of the Landauer and Landauer-Bennett limits—with experimental data. We specifically consider the results of four recent experiments, the only experiments of which we are aware (at this printing) that have attempted to probe the energetics of Landauer erasure with sub-$k_B T$-level resolution.

We preface this discussion by noting that, strictly speaking, experiments cannot be expected to "confirm" the Landauer limit in the usual sense—by establishing numerical agreement between measured values and theoretical predictions—simply because the Landauer limit is a *lower bound* on energy dissipation that does not automatically carry a claim of achievability in principle or in practice. An experimental result can thus either lie within a range of values that are consistent with this lower bound or within a range of values that violate it. Having said that, we also note that the results of these recent experiments not only support the unconditional erasure bound (12) and the conditional erasure bound (3)—generalizations of the Landauer and Landauer-Bennett limits respectively—but also the asymptotic achievability of both. We discuss these experiments and their connection to the present work below.

Three of these experiments, which we discuss first, probed unconditional Landauer erasure in experimental realizations of the model one-bit memory system used in Landauer's original work; a symmetric double well potential. The system "occupies" one of the two potential minima to store a bit value, the stability of which is ensured by a sufficiently large potential barrier that must be surmounted to switch between minima, and is erased by slowly reducing the barrier, tilting the potential to localize the system in the potential minimum designated as the "reset" state, restoring the barrier, and removing the potential tilt. The nature of the information-bearing system, the controlled variation of the double-well potential required to execute the erasure protocol, and measurement of dissipated energy all differ in the three experiments.

In the first unconditional erasure experiment of Berut and co-workers [26], the system is a silica bead suspended in water and manipulated by optical tweezers. An effective double-well potential is created by a laser beam alternately focused at two nearby points, the intensity and positioning of which is varied as required for implementation of the erasure protocol. The position of the particle is tracked in time, and the work done on the particle—all of which is assumed dissipated into the bath—is obtained by time integration of the bead velocity and the potential gradient. The second experiment, due to Jun and co-workers [27], is similar in many ways to that of Berut et al., although the system is a florescent particle in colloidal suspension and the desired potential is created and manipulated by application and variation of a particle-position-dependent electrostatic force. Work done in the erasure operation is inferred from the particle position and potential history throughout the cycle. Finally, in the experiment of Hong and co-workers [28], the system is a single-domain nanomagnet. The double-well potential is built

in to this bistable system, and the barrier and tilt are manipulated via applied magnetic fields oriented along the "hard" and "easy" nanomagnet axes, respectively. Energy dissipation was inferred from hysteresis in the nanomagnet magnetization as measured by the magneto-optical Kerr effect.

In all three unconditional erasure experiments, results were obtained for erasure cycles that started with the respective memory systems in known encoding states but that employed erasure protocols that did not make use of this knowledge of the initial encoding state or copies of the encoding state: they clearly correspond to unconditional Landauer erasure of known data.[15] Furthermore, in all three experiments a single value for the energy dissipation associated with erasure was obtained by averaging over multiple erasure cycles originating in each of the two possible initial states, so they correspond to the same kind of average (2) that is lower bounded by the unconditional erasure bound (12). Finally, all three experiments were designed so the self entropies of the two data states would be identical to one another, and also to the reset state since it is identical to one of the data states, so the first term in the unconditional erasure bound reflecting self entropy changes on individual trials vanishes (12). For erasure of one bit ($\Delta I_{\text{er}} = -1$), the unconditional erasure bound specialized to all three experimental scenarios is thus simply $\langle \Delta \langle E_i^{\mathcal{E}} \rangle \rangle \geq k_B T \ln(2)$.

For slow unconditional erasure processes, where one would expect Landauer's bound to be asymptotically achievable if it is achievable at all, Berut and co-workers report dissipation values that, to within experimental error, saturate in the $k_B \ln(2) - k_B T$ range (depending on erasure success probability) [26]. Jun and co-workers, who were able to achieve Landauer erasure with essentially unit success, also studied dissipation as a function of erasure time, and report clear approach to an asymptotic value of $(0.71 \pm 0.03) k_B T$. Finally, Hong and co-workers report an erasure dissipation of $(1.45 \pm 0.35) k_B T$ for their experiment. These results, taken at face value, support the validity of Landauer's bound for unconditional erasure—and its achievability—in experiments that are designed to have sufficiently high resolution to detect violations.

The fourth experiment, due to Orlov and co-workers [29], is unique in that it probes reduction in erasure cost available through conditioning. In this experiment, the information-bearing system is simply a capacitor in series with a resistor that is prepared in one of two symmetric (equal but opposite polarity) charge states to store a bit value. In contrast to the three experiments discussed above, the reset state is an additional "null" state that does *not* coincide with either of the two "data" states; here it is the "discharged" state of the capacitor. Erasure thus of a stored bit thus corresponds to a discharging of the capacitor, bringing its terminal voltage from some "bit voltage" $V = +V_s$ or $-V_s$ to $V = 0$. This can be done unconditionally or conditionally to achieve unconditional or conditional erasure.

[15] Recall from Sect. 3.3 that, with the way we have defined unknown data in this work, unconditional erasure of both known and unknown data carry the Landauer cost.

Orlov et al. achieve unconditional erasure in this experiment simply by grounding the "hot" terminal of the capacitor, allowing the capacitor to discharge freely through the resistor, which involves no conditioning on the initial capacitor state. They achieve conditional erasure by ramping the terminal voltage from its initial value, which could be either $V = +V_s$ or $-V_s$, to $V = 0$. This is specifically conditional erasure of known data; selection of the appropriate ramping protocol ($+V_s$ to $V = 0$ or $-V_s$ to $V = 0$) is conditioned on the initial charge state of the capacitor. The amount of energy dissipated to the environment during Landauer erasure via Joule heating of the resistor is obtained from measurement of the resistor voltage throughout unconditional and conditional erasure cycles. The experiments were performed at $T = 300\,\mathrm{K}$, and the charge states for bit encoding were selected so the stored energy is $30k_BT$ (i.e. far exceeding the thermal energy).

In the unconditional erasure operation, all of the energy initially stored in the capacitor is, of course, dissipated in the resistor, and this energy far exceeds the Landauer limit. In the conditional erasure operation, however, Orlov et al. measured energy dissipation values as low as $0.01k_BT$ for the lowest ramp-discharge rates they examined. This provides direct comparison of energy dissipation to the environment for unconditional and conditional operations—using the same system and the same initial and final states for both cases—and a demonstration of local energy dissipation below the Landauer limit in the conditional erasure case. While direct application of the bounds (3) and (12) to this system is both questionable and infeasible, both because the capacitor is an open system that exchanges particles with an external source/sink and because the self entropies of the charged encoding states and the null reset state are unknown *and* differ from one another by an unknown amount, the results for energy dissipation in conditional and unconditional erasure straddle $k_BT\ln(2)$ and differ by more than $k_BT\ln(2)$. They are in this sense consistent with expectations from the Landauer and Landauer-Bennett limits. This experiment also provides a particularly clear illustration of the distinction between the energy stored in the states that encode data and the amount of energy dissipated in the erasure of this data.

6 Summary and Conclusion

The Landauer and Landauer-Bennett limits express fundamental lower bounds on the physical costs of unconditional and conditional information erasure, and are some of the earliest and most durable links to have been forged between information and physics. As such, they are cornerstone results in the ongoing project of identifying the essential physical dimensions of information processing and their consequences. At present—half a century since their introduction and perhaps a decade or so before their implications steer the evolution of information technology—these results remain controversial. Some of the controversy is rooted in substantive conceptual and methodological concerns, such as the physical characterization of conditioning, the connection between conditioning of erasure

operations and erasure costs, the relationships between information and various forms of entropy in physical contexts, and the assignment of probabilities in the application of thermodynamics to the erasure problem. Others are by-products of insufficient precision in making and interpreting claims—as when claims of Landauer's limit as inviolable are interpreted as claims that it is achievable—and in defining the key systems and quantities. Yet others stem from lack of a common understanding of crucial distinctions, such as those between physical states that encode known data, random data, unknown data, and no data at all. Relitigation of controversial issues in familiar terms has not yielded resolution.

In this chapter, we presented new proofs of dissipation bounds that generalize the Landauer and Landauer-Bennett limits. These proofs, which are based on fundamental physical descriptions of conditional and unconditional of Landauer erasure, lower bound the erasure-induced system-to-environment energy transfer from lawful quantum dynamics and entropic inequalities alone. They are specifically constructed so they sidestep methodological objections to familiar approaches, most of which are related to demonstrations based on classical thermodynamics and the specific model systems used in these demonstrations. We began by defining and describing Landauer erasure in very general physical terms in Sect. 2. In Sect. 3, we then proved conditional and unconditional erasure bounds that are based on this description, and discussed interpretive issues concerning conditioning, the physical meaning of encoding data, and the role of system boundaries. Next, we showed in Sect. 4 that many objections to previous demonstrations and proofs of the Landauer and Landauer-Bennett limits are sidestepped in our proofs and we discussed their proper interpretation. Finally, in Sect. 5, we showed that the results of recent experiments that probe Landauer erasure with high energetic resolution are consistent with the bounds obtained here. Our theoretical results support the Landauer and Landauer-Bennett limits as immediate and transparent consequences of physical law. They do not carry a claim that these limits are achievable. They are, however, consistent with available experimental results that, taken at face value, *do* support their near achievability.

We hope to have demonstrated that considerable clarity on issues surrounding the energy cost of information erasure can be achieved through explicit physical grounding the relevant notions such as those related to conditioning. In the present work, however, we stopped just short of a thoroughly physical description of information erasure, in that information itself—that which is being erased—was not defined physically. This was deliberate, since our objective here was to provide a rigorous alternative to existing proofs of the Landauer and Landauer-Bennett limits that sidesteps objections leveled at other proofs and demonstrations without departing from them too drastically. Yet, in following the common practice associating information content with the Shannon entropy of the encoding probabilities, which emerged in the proof of our unconditional erasure bound, we have left the physical status of information unresolved. We should hope to do better, not least because the unresolved physical status of information is perhaps the deepest source of confusion surrounding Landauer's limit. How, after all, can we unambiguously tally the physical cost of something that is not unambiguously physical?

In related work, we have aimed to remedy this situation. Using an approach that has much in common with that of the present work, but that formalizes information as an explicitly physical quantity, we have obtained precisely the same conditional and unconditional erasure bounds of this work for the same physical erasure scenarios [36]. The relevant physical conception of information—observer-local referential (OLR) information—distinguishes physical states that bear information from physical states that do not (but are statistically identical), allows information to be clearly distinguished from entropy, places information on an equal footing with other physical state quantities, and defines information in a manner that comports with common conceptions of information in computation [35]. It differs fundamentally from Shannon's mathematical entropy measure in multiple respects, but coincides with it in the scenarios discussed in this work.

In addition to resolving the physical status of information, use of OLR information allows fundamental physical descriptions of information processing to be substantially generalized to naturally accommodate classical and quantum indistinguishability of information bearing states and noisy and otherwise imperfect operations. Furthermore, it enables isolation and tracking of information and lower bounding the dissipative consequences of irreversible information loss in information erasure [36, 42, 43], overwriting [44], implementation of logical transformations [45], and more complex unconditional and conditional computational operations executed in concrete computing circuits [46, 47] and other computing structures [48–50]. By taking this further step of giving information a secure physical grounding, the approach of the present work is greatly generalized and can systematically be applied to a vastly wider range of information processing scenarios. We suspect that exploration of the connection between the present approach and this more general approach can shed additional light on the physical origins of the Landauer and Landauer-Bennett limits.

Appendix

Here, for convenience, we catalog several established properties of von Neumann entropy, unitary transformations, and trace operations that have been used in this work.

1. *von Neumann Entropy is Subadditive:* For any state $\hat{\rho}^{\mathcal{SE}}$,

$$S(\hat{\rho}^{\mathcal{SE}}) \leq S(\hat{\rho}^{\mathcal{S}}) + S(\hat{\rho}^{\mathcal{E}})$$

where

$$\hat{\rho}^{\mathcal{S}} = Tr_{\mathcal{E}}[\hat{\rho}^{\mathcal{SE}}] \qquad \hat{\rho}^{\mathcal{E}} = Tr_{\mathcal{S}}[\hat{\rho}^{\mathcal{SE}}].$$

Equality is achieved in this bound when $\hat{\rho}^{\mathcal{SE}}$ is separable ($\hat{\rho}^{\mathcal{SE}} = \hat{\rho}^{\mathcal{S}} \otimes \hat{\rho}^{\mathcal{E}}$).

2. *Global von Neumann Entropy is Invariant under Unitary-Similarity Transformations:* For any state $\hat{\rho}^{\mathcal{SE}}$ and any unitary operator $\hat{U}$,

$$S(\hat{\rho}^{\mathcal{SE}'}) = \hat{U}\,\hat{\rho}^{\mathcal{SE}}\,\hat{U}^{\dagger} = S(\hat{\rho}^{\mathcal{SE}}).$$

3. *Partovi's Inequality:* For unitary evolution

$$\hat{\rho}^{\mathcal{SE}'} = \hat{U}(\rho^{\mathcal{S}} \otimes \hat{\rho}^{\mathcal{E}}_{\text{th}})\hat{U}^{\dagger}$$

of a system initially in *any* state $\hat{\rho}^{\mathcal{S}}$ and an environment initially in a thermal state $\hat{\rho}^{\mathcal{S}}_{\text{th}}$ at temperature T, Partovi [32] showed that[16]

$$\Delta\langle E^{\mathcal{E}}\rangle \geq k_B T \ln(2) \Delta S^{\mathcal{E}}$$

where

$$\Delta\langle E^{\mathcal{E}}\rangle = \langle E^{\mathcal{E}'}\rangle - \langle E^{\mathcal{E}}\rangle = Tr[\hat{\rho}^{\mathcal{E}'}\hat{H}^{\mathcal{E}}_{\text{self}}] - Tr[\hat{\rho}^{\mathcal{E}}\hat{H}^{\mathcal{E}}_{\text{self}}].$$

4. *Linearity of Unitary-Similarity Transformations:* For a unitary operator $\hat{U}$ and sum $\sum_i p_i\hat{\rho}_i$ of operators $\hat{\rho}_i$,

$$\hat{U}\left(\sum_i p_i\hat{\rho}_i\right)\hat{U}^{\dagger} = \sum_i p_i(\hat{U}\,\hat{\rho}_i\,\hat{U}^{\dagger}).$$

5. *Grouping Property of von Neumann Entropy:* For a set $\hat{\rho}_i$ of density operators with support on orthogonal subspaces, the von Neumann entropy of the convex combination

$$\hat{\rho} = \sum_i p_i\hat{\rho}_i$$

is[17]

$$S(\hat{\rho}) = H(\{p_i\}) + \sum_i p_i S_i(\hat{\rho})$$

[16] This inequality appears in [32] as $\Delta(S_b - \beta U_b) \leq 0$, where ΔS_b, ΔU_b, and β are denoted here as $\Delta S^{\mathcal{E}}$, $\Delta\langle E^{\mathcal{E}}\rangle$ and $(k_B T)^{-1}$ but carry the same meanings. The factor of ln(2) accounts for the differences in the base of the logarithm used to define von Neumann entropy by Partovi and ourselves; Partovi's inequality is based on the thermodynamic definition $S(\hat{\rho}) = -Tr[\hat{\rho}\ln\hat{\rho}]$, which we have reexpressed here in terms of the information-theoretic definition $S(\hat{\rho}) = -Tr[\hat{\rho}\log_2\hat{\rho}]$.

[17] See, for example, Theorem 11.8 of Ref. [51].

where

$$H(\{p_i\}) = -\sum_i p_i \log_2 p_i.$$

6. *Unitary Evolution Preserves Orthogonality:* Consider a unitary $\hat{U}$ and two density operators $\hat{\rho}_i^{\mathcal{SE}}$ and $\hat{\rho}_{i'}^{\mathcal{SE}}$. If $\hat{\rho}_i^{\mathcal{SE}}$ and $\hat{\rho}_{i'}^{\mathcal{SE}}$ are orthogonal, i.e. if

$$\hat{\rho}_i^{\mathcal{SE}} \hat{\rho}_{i'}^{\mathcal{SE}} = 0$$

then

$$\hat{\rho}_i^{\mathcal{SE}'} \hat{\rho}_{i'}^{\mathcal{SE}'} = 0$$

where

$$\hat{\rho}_i^{\mathcal{SE}'} = \hat{U} \hat{\rho}_i^{\mathcal{SE}} \hat{U}^\dagger \qquad \hat{\rho}_{i'}^{\mathcal{SE}'} = \hat{U} \hat{\rho}_{i'}^{\mathcal{SE}} \hat{U}^\dagger.$$

Note that this global preservation of orthogonality on $\mathcal{SE}$ does not imply local preservation of orthogonality on $\mathcal{S}$ and/or $\mathcal{E}$.

References

1. R. Landauer, Irreversibility and heat generation in the computing process. IBM J. Res. Dev. **5**, 183–191 (1961)
2. C.H. Bennett, Logical reversibility of computation. IBM J. Res. Dev. **17**, 525–532 (1973)
3. C.H. Bennett, The thermodynamics of computation—a review. Int. J. Theor. Phys. **21**, 905–940 (1982)
4. W. Porod, R.O. Grondin, D.K. Ferry, G. Porod, Dissipation in computation. Phys. Rev. Lett. **52**, 232 (1984)
5. P. Benioff, Comment on "Dissipation in computation". Phys. Rev. Lett. **53**, 1203 (1984)
6. T. Toffoli, Comment on "Dissipation in computation". Phys. Rev. Lett. **53**, 1204 (1984)
7. R. Landauer, Dissipation in computation. Phys. Rev. Lett. **53**, 1205 (1984)
8. W. Porod, R.O. Grondin, D.K. Ferry, G. Porod, Porod *et al.* respond. Phys. Rev. Lett. **53**, 1206 (1984)
9. C.H. Bennett, Notes on the history of reversible computation. IBM J. Res. Dev. **32**, 16–23 (1988)
10. J. Earman, J.D. Norton, Exorcist XIV: the wrath of Maxwell's demon. Part II. From Szilard to Landauer and beyond. Stud. Hist. Philos. Sci. B: Stud. Hist. Philos. Mod. Phys. **30**, 1–40 (1999)
11. C.H. Bennett, Notes on Landauer's principle, reversible computation, and Maxwell's demon. Stud. Hist. Philos. Sci. B: Stud. Hist. Philos. Mod. Phys. **34**, 501–510 (2003)
12. J.D. Norton, Eaters of the lotus: Landauer's principle and the return of Maxwell's demon. Stud. Hist. Philos. Sci. B: Stud. Hist. Philos. Mod. Phys. **36**, 375–411 (2005)
13. C.S. Lent, M. Liu, Y. Lu, Bennett clocking of quantum-dot cellular automata and the limits to binary logic scaling. Nanotechnology **17**, 4240 (2006)

14. V.V. Zhirnov, R.K. Cavin, Comment on "Bennett clocking of quantum-dot cellular automata and the limits to binary logic scaling". Nanotechnology **18**, 298001 (2007)
15. C.S. Lent, Reply to "Comment on 'Bennett clocking of quantum-dot cellular automata and the limits to binary logic scaling' ". Nanotechnology **18**, 298002 (2007)
16. J.D. Norton, Waiting for Landauer. Stud. Hist. Philos. Sci. B: Stud. Hist. Philos. Mod. Phys. **42**, 184–198 (2011)
17. J. Ladyman, K. Robertson, Landauer defended: reply to Norton. Stud. Hist. Philos. Sci. B: Stud. Hist. Philos. Mod. Phys. **44**, 263–271 (2013)
18. J.D. Norton, Author's reply to Landauer defended. Stud. Hist. Philos. Mod. Phys. **44**, 272–272 (2013)
19. J. Ladyman, K. Robertson, Going round in circles: Landauer vs. Norton on the thermodynamics of computation. Entropy **16**, 2278–2290 (2014)
20. G.P. Boechler, J.M. Whitney, C.S. Lent, A.O. Orlov, G.L. Snider, Fundamental limits of energy dissipation in charge-based computing. Appl. Phys. Lett. **97**, 103502 (2010)
21. V.V. Zhirnov, R.K. Cavin, Comment on "Fundamental limits of energy dissipation in charge-based computing". Appl. Phys. Lett. **97**, 103502 (2010)
22. G. Boechler, J. Whitney, C. Lent, A. Orlov, G. Snider, Response to "Comment on 'Fundamental limits of energy dissipation in charge-based computing"'. Appl. Phys. Lett. **98**, 096101 (2011)
23. L.B. Kish, C.G. Granqvist, S.P. Khatri, F. Peper, Zero and negative energy dissipation at information-theoretic erasure. J. Comput. Electron. **15**, 335–339 (2016)
24. N.G. Anderson, Comment on "Zero and negative energy dissipation at information-theoretic erasure". J. Comput. Electron. **15**, 340–342 (2016)
25. L.B. Kish, C.G. Granqvist, S.P. Khatri, F. Peper, Response to "Comment on 'Zero and negative energy dissipation at information-theoretic erasure"'. J. Comput. Electron. **15**, 343–346 (2016)
26. A. Berut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, E. Lutz, Experimental verification of Landauer's principle linking information and thermodynamics. Nature **483**, 187–189 (2012)
27. Y. Jun, M. Gavrilov, J. Bechhoefer, High-precision test of Landauer's principle in a feedback trap. Phys. Rev. Lett. **113**, 190601 (2014)
28. J. Hong, B. Lambson, S. Dhuey, J. Bokor, Experimental test of Landauer's principle in single-bit operations on nanomagnetic memory bits. Sci. Adv. **2**, 1501492 (2016)
29. A.O. Orlov, C.S. Lent, C.C. Thorpe, G.P. Boechler, G.L. Snider, Experimental test of Landauer's principle at the sub-$k_B T$ level. Jpn. J. Appl. Phys. **51**, 06FE10 (2012)
30. T.N. Theis, H.S.P. Wong, The end of Moore's law: a new beginning for information technology. Comput. Sci. Eng. **19**, 41–50 (2017)
31. M.P. Frank, Throwing computing into reverse. IEEE Spectr. **54**, 32–37 (2017)
32. M.H. Partovi, Quantum thermodynamics. Phys. Lett. A **137**, 440–444 (1989)
33. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
34. C.H. Bennett, R. Landauer, The fundamental physical limits of computation. Sci. Am. **253**, 48–57 (1985)
35. N.G. Anderson, Information as a physical quantity. Inform. Sci. **415–416**, 397–413 (2017)
36. N.G. Anderson, Landauer's limit and the physicality of information. Eur. Phys. J. B. **91**, 156–164 (2018)
37. M. Lopez-Suarez, I. Neri, L. Gammaitoni, Sub-$k_B T$ micro-electromechanical irreversible logic gate. Nat. Commun. **7**, 12068 (2016)
38. J. Ladyman, What does it mean to say that a physical system implements a computation? Theor. Comput. Sci. **410**, 376–383 (2009)
39. H.S. Leff, A.F. Rex, Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing (Institute of Physics, Bristol, 2003)
40. L.B. Kish, D.K. Ferry, Information entropy and thermal entropy: apples and oranges. J. Comput. Electron. **17**, 43–50 (2018)
41. J. Ladyman, S. Presnell, A.J. Short, B. Groisman, The connection between logical and thermodynamic irreversibility. Stud. Hist. Philos. Mod. Phys. **38**, 58–79 (2006)

42. N.G. Anderson, Information erasure in quantum systems. Phys. Lett. A **372**, 5552–5555 (2008)
43. N.G. Anderson, Irreversible information loss: fundamental notions and entropy costs. Int. J. Mod. Phys.: Conf. Ser. **33**, 1460354 (2014)
44. N.G. Anderson, Overwriting information: correlations, physical costs, and environment models. Phys. Lett. A **376**, 1426–1433 (2012)
45. N.G. Anderson, On the physical implementation of logical transformations: generalized L-machines. Theor. Comput. Sci. **411**, 4179–4199 (2010)
46. I. Ercan, N.G. Anderson, Heat dissipation in nanocomputing: lower bounds from physical information theory. IEEE Trans. Nanotechnol. **12**, 1047–1060 (2013)
47. K.J. Stearns, N.G. Anderson, Throughput-dissipation tradeoff in partially reversible nanocomputing: a case study. In: *Proceedings of 2013 IEEE/ACM International Symposium on Nanoscale Architectures* (IEEE Press, New York, 2013), pp. 101–105
48. N. Ganesh, N.G. Anderson, Irreversibility and dissipation in finite-state automata. Phys. Lett. A **377**, 3266–3271 (2013)
49. N.G. Anderson, I. Ercan, N. Ganesh, Toward nanoprocessor thermodynamics. IEEE Trans. Nanotechnol. **12**, 902–909 (2013)
50. J. Ricci, N.G. Anderson, Architecture and dissipation: on the energy costs of general purposeness in von Neumann processors. In: *Proceedings of the 2017 IEEE International Conference on Rebooting Computing* (IEEE Press, New York, 2017), pp. 194–198
51. M.A. Nielsen, I.L. Chuang, Quantum Computation and Quantum Information (Cambridge University Press, Cambridge, 2000)

Second Law, Entropy Production, and Reversibility in Thermodynamics of Information

Takahiro Sagawa

Contents

1 Introduction

Thermodynamics is intrinsically related to information, as the entropy represents the lack of our knowledge. The first clue on the information-thermodynamics link was provided by Maxwell, who considered a thought experiment of "Maxwell's demon" [1]. Later, Szilard suggested a quantitative connection between work extraction and information gain [2], even several decades before the establishment of information theory by Shannon [3]. The role of information in thermodynamics was investigated and controversies were raised throughout the twentieth century [4–10].

In this decade, thermodynamics of information has attracted renewed attention because of the development of the modern theory of nonequilibrium thermodynamics [11–23], which is often referred to as stochastic thermodynamics. Especially, a fundamental thermodynamic relation called the fluctuation theorem was discovered in 1990s [11, 12, 14, 15], which has opened up a new avenue of research. We note, however, that only a few seminal works have been done already in the 1970s and 80s [24–28].

T. Sagawa (✉)
Department of Applied Physics, The University of Tokyo, Tokyo, Japan
e-mail: sagawa@ap.t.u-tokyo.ac.jp

C. S. Lent et al. (eds.), *Energy Limits in Computation*,
https://doi.org/10.1007/978-3-319-93458-7_3

Thermodynamics of information can now be formulated based on stochastic thermodynamics [29], by incorporating information concepts such as Shannon entropy and mutual information. Specifically, the Landauer principle for information erasure [30–37] and feedback control by Maxwell's demon [38–51] have been investigated from the modern point of view.

Furthermore, the relationship between thermodynamics and information has become significant from the experimental point of view [52]. The first quantitative demonstration of the work extraction by Maxwell's demon was performed in Ref. [53], and the Landauer bound for information erasure was demonstrated in Ref. [54]. Several fundamental experiments have been further performed both in the classical and quantum regimes [55–69].

In this article, we review the theoretical foundation of thermodynamics of information. Especially, we aim at clarifying the concept of reversibilities and the consistency between Maxwell's and the second law, which we hope would unravel subtle conceptual problems.

We will put special emphasis on the following two observations. First, reversibility has several different aspects. In particular, thermodynamic reversibility and logical reversibility are fundamentally distinct concepts, which are associated with different kinds of the degrees of freedom of a thermodynamic system. Second, mutual information is a crucial concept to understand the consistency between the demon and the second law. We will see that the demon is consistent with the second law for the measurement and the feedback processes *individually*.

We here make some side remarks. First, in this article, we only consider classical systems in the presence of an infinitely large heat bath, though essentially the same argument applies to quantum systems. Second, this article is completely newly written, but is closely related to a paper [37] by the author, where the present article is intended to be more pedagogical and comprehensive. Finally, for simplicity of notation, we set the Boltzmann constant to unity (i.e., $k_{\rm B} = 1$) throughout the article.

This article is organized as follows. In the rest of this section, we briefly summarize the above-mentioned two observations. In Sect. 2, we review the second law and reversibility in conventional thermodynamics as a preliminary. In Sect. 3, we discuss the framework of stochastic thermodynamics and clarify what reversibility means there. In Sect. 4, we review reversibility in computation, which is referred to as logical reversibility. In Sect. 5, we discuss thermodynamics of information in a simple setup, and state the Landauer principle. The relationship between thermodynamic reversibility and logical reversibility is clarified in this simple setup. In Sect. 6, we generally formulate thermodynamics of computation. In Sect. 7, we slightly change the topic and discuss work extraction by Maxwell's demon. In particular, we consider the upper bound of extractable work and formulate thermodynamic reversibility with feedback. In Sect. 8, we generally discuss the entropy balance during the measurement and the feedback processes of the demon and clarify how the demon is consistent with the second law in these processes. In Sect. 9, we make concluding remarks, where we briefly summarize some topics that are not mentioned in the preceding sections.

◊◊◊

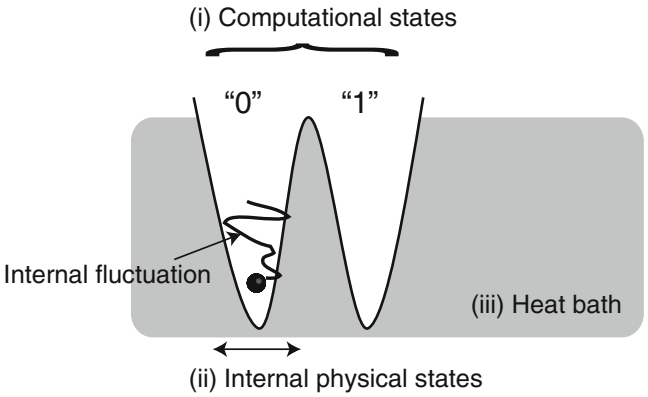

Fig. 1 A schematic of a binary memory, which is modeled as a Brownian particle in a double-well potential. The memory represents computational state "0" or "1", when the particle is in the left or right well, respectively. These wells are separated by a barrier that is sufficiently higher than thermal fluctuations. The internal physical degrees of freedom represent the position of the particle inside individual wells, where the particle suffers thermal fluctuations inside these wells. The entire memory is attached to a heat bath that is in thermal equilibrium

At this stage, let us summarize some key observations, which will be detailed in the following sections. We consider a binary memory that stores one bit of information ("0" and "1"), and suppose that the memory is in contact with a single heat bath (see Fig. 1). We focus on the relationship among thermodynamic reversibility, logical reversibility, and heat emission from the memory. We first note that there are three kinds of the degrees of freedom in this setup:

(i) The computational states of the memory (i.e., "0" and "1" for the binary case). Such computational states should be robust against thermal fluctuations, in order to store information stably.
(ii) Internal physical states of the memory, which represent the physical degrees of freedom associated with a single computational state.
(iii) The degrees of freedom of the heat bath, which is assumed to be in thermal equilibrium.

Then, we have the following observations:

- Thermodynamic reversibility refers to the reversibility of the total system including the heat bath and thus is connected to the entropy change in (i)+(ii)+(iii).
- Logical reversibility refers to the reversibility of the computational states only and thus is connected to the entropy change in (i).
- Heat transfer to the bath is bounded by the entropy change of all the degrees of freedom of the memory, i.e., (i)+(ii).

Therefore, the above three concepts should be distinguished fundamentally, while some of them can become equivalent in some specific setups.

Entropy production characterizes thermodynamic reversibility, and thus is related to the degrees of freedom (i)+(ii)+(iii). A general version of the second law of thermodynamics states that entropy production is always nonnegative for any transition from a nonequilibrium state to another nonequilibrium state (see Sect. 3 for details). In particular, the entropy production of the total system, including an engine and the memory of Maxwell's demon, is nonnegative for individual processes of measurement and feedback control. A crucial point here is that the mutual information between the engine and the demon should be counted as a part of the entropy production. By doing so, the demon is always consistent with the second law of thermodynamics, and we do not need to consider the information-erasure process to understand the consistency.

2 Reversibility in Conventional Thermodynamics

As a preliminary, we briefly review conventional thermodynamics, which has been established in the nineteenth century as a phenomenological theory for macroscopic systems. A remarkable feature of conventional thermodynamics lies in the fact that it can be formulated in a self-contained manner, without referring to underlying microscopic dynamics such as Newtonian mechanics and quantum mechanics. In fact, an equilibrium state is characterized by only a few macroscopic quantities such as the energy and the temperature. We can also define the thermodynamic entropy in a purely phenomenological manner by using, for example, the Clausius formula [70].

We note that conventional thermodynamics can be formulated as a mathematically rigorous axiomatic theory [71]. While in this article we do not formalize our argument in a rigorous manner, the following argument of this section can be made rigorous in line with the theory of Lieb and Yngvason [71].

We focus on the situation that a thermodynamic system is in contact with a single heat bath at temperature T. Let $\beta := T^{-1}$ be the inverse temperature. We consider a transition from an equilibrium state to another equilibrium state. During the transition, the system absorbs the heat Q from the bath, and changes its thermodynamic entropy by ΔS_{T}. We note that in conventional thermodynamics, the thermodynamic entropy S_{T} is defined only for equilibrium states, and the second law only concerns a transition from an equilibrium state to another equilibrium state, though intermediate states can be out of equilibrium.

Then, the second law is stated as follows.

Second Law of Conventional Thermodynamics *An equilibrium state can be converted into another equilibrium state with heat absorption Q, if and only if*

$$\Delta S_{\mathrm{T}} - \beta Q \geq 0. \tag{1}$$

We note that the "only if" part (i.e., any possible state conversion satisfies inequality (1)) is the usual second law, while "if" part (i.e., a state conversion is possible if inequality (1) is satisfied) is also true under reasonable axioms [71].

The left-hand side of the second law (1) is referred to as the entropy production, which we denote by

$$\Sigma := \Delta S_{\mathrm{T}} - \beta Q. \tag{2}$$

This terminology, the entropy production, dates back to Prigogine, who associated $-\beta Q$ with the entropy change of the bath [72]. In this spirit, Σ is regarded as the entropy change of the entire "universe" that consists of the system and the heat bath.

We can also rewrite the second law (1) in terms of the work and the free energy. Let W be the work performed on the system, and ΔE be the change in the average internal energy. The first law of thermodynamics is given by

$$W + Q = \Delta E. \tag{3}$$

By substituting the first law into inequality (1), we obtain

$$W \geq \Delta F_{\mathrm{eq}}, \tag{4}$$

where

$$F_{\mathrm{eq}} := E - T S_{\mathrm{T}} \tag{5}$$

is the equilibrium free energy. If the process is cyclic, inequality (4) reduces to $W \geq 0$, which is the Kelvin's principle, stating that perpetual motion of the second kind is impossible (i.e., a positive amount of work cannot be extracted from an isothermal cycle).

◊◊◊

We next formulate the concept of reversibility in conventional thermodynamics. Based on the standard textbook argument [70], we adopt the following definition:

Definition (Reversibility in Conventional Thermodynamics) A state transition from one to another equilibrium state is thermodynamically reversible, if and only if the final state can be restored to the initial state, without remaining any effect on the outside world.

We note that "effect" above is regarded as a "macroscopic effect" in conventional thermodynamics because microscopic changes (i.e., the subleading terms in the thermodynamic limit) are usually neglected.

A crucial feature of this definition is that thermodynamic reversibility is completely characterized by the entropy production, as represented by the following theorem.

Theorem *Thermodynamic reversibility is achieved if and only if the entropy production is zero, i.e.,* $\Sigma = 0$.

Proof While one can find a proof of the above theorem in standard textbooks of thermodynamics (at least implicitly), we reproduce it here for the sake of self-containedness.

i) Suppose that a transition is thermodynamically reversible. Then there exists a reverse transition that satisfies the requirements in the definition of reversibility. From the requirement that there is no remaining effect in the outside world, $Q_{\rm reverse} = -Q$ should hold, because otherwise the energy of the heat bath is changed after the reverse transition. Combining this with $\Delta S_{\rm T,reverse} = -\Delta S_{\rm T}$, we have $\Sigma_{\rm reverse} = -\Sigma$. On the other hand, both of $\Sigma \geq 0$ and $\Sigma_{\rm reverse} \geq 0$ hold from the second law. Therefore, $\Sigma = \Sigma_{\rm reverse} = 0$.
ii) Suppose that Σ is zero. Then $-\Sigma = (-\Delta S_{\rm T}) - \beta(-Q)$ is also zero. Therefore, the reverse transition is possible with $Q_{\rm reverse} := -Q$, because of the "if" part of the second law. □

We consider the concept of quasi-static process, by which we define that the system remains in equilibrium during the entire process. Thermodynamic reversibility is achieved if a process is quasi-static, because the quasi-static condition guarantees that $\Delta S_{\rm T} \simeq \beta Q$. The quasi-static limit is achieved by an infinitely slow process in many situations in which

$$\Delta S_{\rm T} = \beta Q + O(\tau^{-1}) \tag{6}$$

holds, where τ is the time interval of the entire process. In the infinitely-slow limit $\tau \to +\infty$, the equality in (1) is achieved. More precisely, $\tau \to +\infty$ means $\tau/\tau_0 \to \infty$, where τ_0 is the relaxation time of the system.

We note, however, that there is a subtle difference between quasi-static and infinitely slow processes in general. For example, suppose that there is a box separated by a wall at the center, and gas is only in the left side of the wall. No matter how slow the removal of the wall is, the free expansion of the gas to the right side is not quasi-static and thermodynamically irreversible. Such a situation has been experimentally demonstrated in Ref. [73] in the context of stochastic thermodynamics.

The foregoing argument can be straightforwardly generalized to situations with multiple heat baths at different temperatures. In particular, the zero entropy production of a heat engine with two baths implies the maximum efficiency of the heat-to-work conversion. For example, the Carnot cycle attains the maximum efficiency, where the entropy production is zero and the cycle is thermodynamically reversible. We note that it has been rigorously proved that any thermodynamically reversible process is infinitely slow, based on some reasonable assumptions (including that fluctuations of the system do not diverge) [74]. Since an infinitely slow process gives the zero power (i.e., the work per unit time is zero), thermodynamically reversible engines might be practically useless. For that reason, the efficiency at the maximum power has been intensively studied [75, 76].

The concept of reversibility in conventional thermodynamics is generalized to stochastic thermodynamics, as discussed in the next section.

3 Reversibility in Stochastic Thermodynamics

Stochastic thermodynamics is an extension of thermodynamics to situations where a system is not necessarily macroscopic, and the initial and final states are not necessarily in thermal equilibrium [21–23]. When a large heat bath is attached to a small system, thermal fluctuations affect the system, and its dynamics become stochastic. Correspondingly, thermodynamic quantities, such as the heat, the work, and the entropy, become stochastic. Biochemical molecular motors and colloidal particles are typical examples of stochastic thermodynamic systems, with which numerous experiments have been performed [52].

Because of thermal fluctuations, the second law of thermodynamics can be violated with a small probability in small systems. At the level of the ensemble average, however, the second law is still valid in an extended form. Furthermore, a universal relation called the fluctuation theorem has been established by taking into account the role of thermal fluctuations of the entropy production, from which the second law of thermodynamics can be reproduced. This is the reason why thermodynamics is still relevant to small systems.

To formulate the second law of stochastic thermodynamics, we need the concept of information entropy, in particular Shannon entropy. Let X be a probability variable, which takes a particular value x with probability $P(x)$. The Shannon entropy of X is then defined as [77]

$$S(X) := -\sum_x P(x) \ln P(x) \geq 0. \tag{7}$$

If the probability variable is continuous, we replace the summation above by an integral over x. We note that if $P(x) = 0$, $P(x) \ln P(x)$ is regarded as zero.

In contrast to the thermodynamic entropy that can be defined only for thermal equilibrium, the Shannon entropy can be defined for an arbitrary probability distribution. However, these entropies coincide in the canonical distribution $P_{\rm can}(x) := e^{\beta(F_{\rm eq}-E(x))}$, where $F_{\rm eq}$ is the equilibrium free energy and $E(x)$ is the Hamiltonian (i.e., the internal energy of state x). In this case, the Shannon entropy is given by the difference between the average energy and the free energy:

$$S(X) := -\sum_x P_{\rm can}(x) \ln P_{\rm can}(x) = \beta \left(\sum_x P_{\rm can}(x) E(x) - F_{\rm eq} \right), \tag{8}$$

which is a statistical-mechanical expression of the thermodynamic entropy.

We suppose that a small system is described by a Markov jump process or overdamped Langevin dynamics [78, 79]. We also suppose that the system is driven by external parameters (e.g., the center and the frequency of optical tweezers), which are represented by time-dependent parameters in a master equation or a Langevin equation. The following argument is independent of the details of dynamics, and therefore we do not explicitly write down stochastic equations.

On the other hand, we assume that variables that break the time-reversal symmetry (i.e., that have the odd parity for the time-reversal transformation) are absent or negligible. In particular, the momentum term is negligible in dynamics of the system (in particular, the Langevin equation is overdamped), and the magnetic field is absent in the external potential. While some of the following arguments are straightforwardly generalized to systems with the momentum and the magnetic field, there are some subtle problems with odd-parity variables.

Let x be the state of the system, which has a certain probability distribution. Correspondingly, we define the Shannon entropy of the system, written as S. We consider a transition from the initial distribution to the final distribution of the system, where both the distributions are arbitrary and can be out of equilibrium. Let ΔS be the change in the Shannon entropy of the system, and Q be the ensemble average of the heat absorption. We then have the following version of the second law:

Second Law of Stochastic Thermodynamics *A distribution can be converted into another distribution with the heat absorption Q, if and only if*

$$\Delta S - \beta Q \geq 0. \tag{9}$$

As in conventional thermodynamics, the left-hand side of inequality (9) is referred to as the (ensemble averaged) entropy production:

$$\Sigma := \Delta S - \beta Q. \tag{10}$$

This is regarded as the total entropy increase in the "whole universe" that consists of the system and the heat bath.

As discussed before, if the distribution is canonical, the Shannon entropy reduces to the thermodynamic entropy. Therefore, inequality (9) is a reasonable generalization of the conventional second law (1) to situations that the initial and final distributions are not necessarily canonical.

We can rewrite inequality (9) in terms of the work and the free energy. Let W be the work performed on the system. We denote the average energy of the system by $E := \sum_x P(x)E(x)$. The first law of thermodynamics is again given by

$$W + Q = \Delta E, \tag{11}$$

where ΔE is the change in the average energy. By substituting (11) into inequality (9), we obtain

$$W \geq \Delta F, \tag{12}$$

where

$$F := E - TS \tag{13}$$

is called the nonequilibrium free energy [13, 20]. We note that if the probability distribution is canonical, we have from Eq. (8) that $F = F_{\rm eq}$. In such a case, inequality (12) reduces to $W \geq \Delta F_{\rm eq}$, which is nothing but the second law of equilibrium thermodynamics.

We now consider the precise meaning of "if and only if" in the second law above. The "only if" part (i.e., any possible transition satisfies inequality (9)) has been proved for Markovian stochastic dynamics, including Markov jump processes and Langevin dynamics [13, 15, 18, 20]. Furthermore, inequality (9) has been proved for a setup in which the total system, including the heat bath, obeys Hamiltonian dynamics [16].

Before discussing the "if" part, we show that there exists a protocol that achieves the equality in (9), for given initial and final distributions and potentials [29]. Such a protocol is schematically shown in Fig. 2, which consists of sudden and infinitely slow changes of the external potential. While the initial distribution can be out of equilibrium, the potential is instantaneously adjusted to the distribution to make

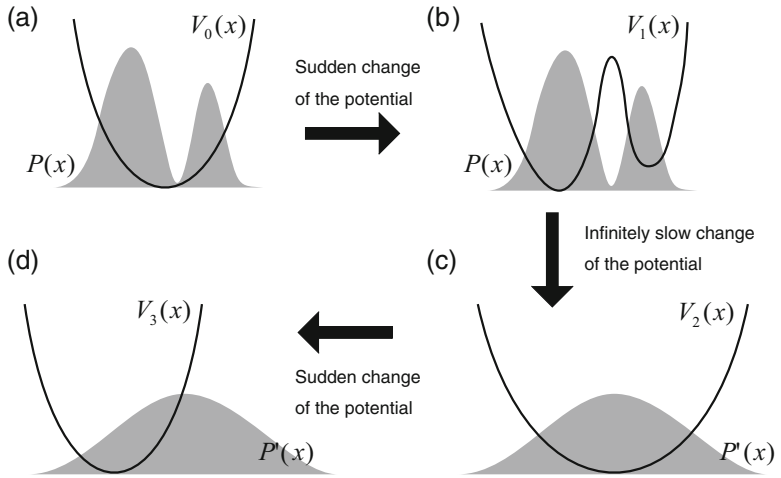

Fig. 2 Protocol that achieves thermodynamic reversibility [29]. (**a**) The probability distribution $P(x)$ (shaded) is in general different from the canonical distribution of the external potential $V_0(x)$ (i.e., the system is out of equilibrium). (**b**) The instantaneous change of the potential from $V_0(x)$ to $V_1(x)$ such that $P(x)$ is now the canonical distribution of $V_1(x)$. (**c**) The potential is infinitely slowly changed from $V_1(x)$ to $V_2(x)$. Correspondingly, the distribution changes infinitely slowly, and ends up with $P'(x)$ that is the canonical distribution of $V_2(x)$. (**d**) The potential is again suddenly changed from $V_2(x)$ to $V_3(x)$. The distribution $P'(x)$ is in general no longer the canonical distribution of $V_3(x)$. During the entire dynamics, the probability distribution does not evolve spontaneously, which makes the entropy production zero

it equilibrium. Then, the potential is changed infinitely slowly, which brings the distribution to the final one. After that, the potential is again changed suddenly, and the final distribution can again be out of equilibrium.

It is easy to check that the entropy production is zero, $\Sigma = 0$, in this protocol. In fact, the heat absorption from (b) to (c) satisfies $\Delta S - \beta Q = 0$, while in the potential switching processes the heat absorption is zero and the Shannon-entropy change is also zero. Therefore, the entropy production of the entire process is zero. The essence of this protocol is that the distribution is always canonical, except for the very moments of the initial and final times. In this sense, this protocol is regarded as quasi-static. In other words, the probability distribution never spontaneously evolves during the entire process, which prohibits the entropy production from becoming positive.

Based on the above protocol, the "if" part of the second law (i.e., a transition is possible if inequality (9) is satisfied) can be shown, by explicitly constructing a protocol with $\Delta S - \beta Q > 0$. For example, we can add an auxiliary cyclic process to an intermediate step of the infinitely slow protocol ((b)–(c) in Fig. 2). If such a cyclic process is not slow, it simply "stirs" the system, and the system emits a positive amount of heat (i.e., $-Q' > 0$). By adding this heat emission, we have a positive amount of entropy production in the entire process.

$\Diamond\Diamond\Diamond$

We next discuss the concept of reversibility in stochastic thermodynamics. While the fundamental idea is the same as in conventional thermodynamics, we need to care about probability distributions in stochastic thermodynamics. We thus adopt the following definition:

Definition (Reversibility in Stochastic Thermodynamics) A stochastic process is thermodynamically reversible, if and only if the final probability distribution can be restored to the initial one, without remaining any effect on the outside world.

As in conventional thermodynamics, reversibility defined above is completely characterized by the entropy production.

Theorem *Reversibility in stochastic thermodynamics is achieved if and only if the entropy production is zero, i.e.,* $\Sigma = 0$.

The proof of this theorem is completely parallel to the case of conventional thermodynamics, just by replacing thermodynamic entropy by Shannon entropy.

From the above theorem, the protocol described in Fig. 2, satisfying $\Sigma = 0$, is thermodynamically reversible. We can also directly see this, because the final distribution is restored to the initial distribution, just by reversing the entire protocol step by step. In this reversed protocol, the heat absorption satisfies $Q_{\text{reverse}} = -Q$, and therefore no effect remains in the outside world.

4 Reversibility in Computation

We next discuss the concept of reversibility in computation, which we will show is fundamentally distinct from thermodynamic reversibility.

Let M be the set of the input states of computation. For example, if any input consists of n binary bits, then $M = \{0, 1\}^n$. We can also consider M' being the set of the output states of computation, which can be different from M in general. Any computation process is a map $\hat{C}$ from M to M'.

We see three simple examples of such computation.

NOT The NOT gate simply flips the input bit: $M = M' = \{0, 1\}$, and

$$\hat{C}(0) = 1, \ \hat{C}(1) = 0. \tag{14}$$

ERASE The information erasure maps any input to a single "standard state." If the input is one bit and the standard state is "0", then $M = M' = \{0, 1\}$, and

$$\hat{C}(0) = 0, \ \hat{C}(1) = 0. \tag{15}$$

AND For the AND gate, the input is two bits and the output is one bit: $M = \{0, 1\}^2$, $M' = \{0, 1\}$, and

$$\hat{C}(00) = 0, \ \hat{C}(01) = 0, \ \hat{C}(10) = 0, \ \hat{C}(11) = 1. \tag{16}$$

Rigorously speaking, a *computable* map from $\mathbb{N}^k$ to $\mathbb{N}$ is defined as a partial recursive function, or equivalently, a partial function that can be implemented by a Turing machine [80]. However, this precise characterization of computability is not necessary for the following argument. We also note that we only consider deterministic computation in this article.

We now define logical reversibility of computation [6, 7, 9]. For a logically reversible computation, one can recover the original input from only the output, which is formalized as follows.

Definition (Logical Reversibility) A deterministic computational process $\hat{C}$ is logically reversible, if and only if it is an injection. In other words, $\hat{C}$ is logically reversible if and only if, for any output, there is a unique input.

In the case of the aforementioned three examples, NOT is logically reversible, while ERASE and AND are logically irreversible. Figure 3 schematically illustrates these examples, where it is visually obvious that only NOT is injection and thus logically reversible.

We next show that logical reversibility can be characterized by the Shannon entropy of the computational states. For that purpose, we consider a probability distribution over inputs. Let $p(m)$ be the probability of input $m \in M$. The probability distribution over the outputs is then given by

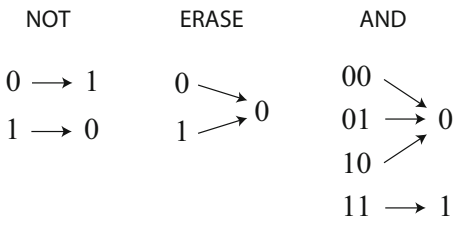

Fig. 3 Three examples of computation. NOT is logically reversible, while ERASE and AND are logically irreversible

$$p(m') = \sum_{m \in \hat{C}^{-1}(m')} p(m), \tag{17}$$

where $m \in \hat{C}^{-1}(m')$ means $m' = \hat{C}(m)$. Correspondingly, we define the Shannon entropies of the input and the output by

$$S(M) := -\sum_{m \in M} p(m) \ln p(m), \quad S(M') := -\sum_{m' \in M'} p(m') \ln p(m'). \tag{18}$$

Then, as a general property of the Shannon entropy [77], we have

$$\Delta S(M) := S(M') - S(M) \leq 0. \tag{19}$$

In fact,

$$\begin{aligned} S(M) - S(M') &= -\sum_{m' \in M'} \sum_{m \in \hat{C}^{-1}(m')} p(m) \ln p(m) + \sum_{m' \in M'} \sum_{m \in \hat{C}^{-1}(m')} p(m) \ln p(m') \\ &= \sum_{m' \in M'} \sum_{m \in \hat{C}^{-1}(m')} p(m) \ln \frac{p(m')}{p(m)} \\ &\geq 0, \end{aligned} \tag{20}$$

where we used Eq. (17) to obtain the second term on the right-hand side of the first line, and used $p(m') \geq p(m)$ with $m' = \hat{C}(m)$ to obtain the last inequality. Therefore, the Shannon entropy does not increase by any deterministic computation.

We show the entropy changes in the aforementioned three examples. We assume that the probability distribution of the input is uniform.

NOT $S(M) = S(M') = \ln 2$, and thus $\Delta S(M) = 0$.
ERASE $S(M) = \ln 2$, $S(M') = 0$, and thus $\Delta S(M) = -\ln 2 < 0$.
AND $S(M) = 2 \ln 2$, $S(M') = -(3/4)\ln(3/4) - (1/4)\ln(1/4)$, and thus $\Delta S(M) = -(3/4)\ln 3 < 0$.

The equality in the last line of (20) is achieved, if and only if $p(m') = p(m)$ holds for any (m, m') satisfying $m' = \hat{C}(m)$ and $p(m) \neq 0$. This is equivalent to the following: For any $m' \in M'$ with $p(m') \neq 0$, there exists a unique $m \in \hat{C}^{-1}(m')$

Table 1 Characterization of thermodynamic and logical reversibilities

	Reversible	Irreversible
Thermodynamically	$\Sigma = 0$	$\Sigma > 0$
Logically	$\Delta S(M) = 0$	$\Delta S(M) < 0$

with $p(m) \neq 0$. This means that $\hat{C}$ is injection, when the domain of $\hat{C}$ is restricted to the set of $m \in M$ with $p(m) \neq 0$. Therefore, we obtain the following theorem (in a slightly rough expression):

Theorem *A deterministic computational process is logically reversible, if and only if the Shannon entropy of the computational process does not change.*

We are now in a position to discuss why logical and thermodynamic reversibilities are fundamentally distinct. In fact, logical reversibility is the reversibility of only computational states (i.e., the degrees of freedom (i) in Sect. 1), and thus characterized by the Shannon-entropy change of computational states, $\Delta S(M)$. On the other hand, thermodynamic reversibility is the reversibility of the entire system including the heat bath (i.e., the degrees of freedom (i)+(ii)+(iii) in Sect. 1), and thus characterized by the total entropy production Σ. This observation is summarized in Table 1. We will further develop this observation in the subsequent sections, especially in the context of the Landauer principle.

We note that any logically irreversible process can be embedded in another logically reversible process by extending the space of computational states [7]. For example, if we prepare $M \times M'$ as an extended set of computational states, we can construct an extended map $\hat{C}'$ by

$$\hat{C}' \ : \ (m, 0) \in M \times M' \ \mapsto \ (m, \hat{C}(m)) \in M \times M', \tag{21}$$

where $0 \in M'$ is the standard state of M'. Strictly speaking, $\hat{C}'$ should be interpreted as a map from $M \times \{0\}$ to $M \times M'$. This extended map $\hat{C}'$ reproduces the original map $\hat{C}$, if we only look at M of the input and M' of the output. A crucial feature of $\hat{C}'$ is that the input $m \in M$ is kept in M of the output of $\hat{C}'$. Therefore, the extended map $\hat{C}'$ is logically reversible, even when the original map $\hat{C}$ is logically irreversible. Such a construction of a logically reversible extension of a logically irreversible map has experimentally be demonstrated in Ref. [64] in the context of thermodynamics of computation.

5 Landauer Principle

We now discuss thermodynamics of computation. Before a general argument, in this section we focus on a paradigmatic model: the conventional setup of the Landauer principle for information erasure [6, 30, 31].

The information erasure is nothing but the ERASE gate discussed in Sect. 4: The initial information of "0" or "1" is erased, so that the final computational state is always in "0" that is called the standard state. This is a logically irreversible process as discussed before.

If the initial distribution of the input is uniform (i.e., $p(m = 0) = p(m = 1) = 1/2$), the Shannon entropy of the input is $S(M) = \ln 2$, while that of the output is $S(M') = 0$. The change in the computational entropy is given by $\Delta S(M) = -\ln 2$, as already shown in Sect. 4. This is the erasure of one bit of information.

To physically implement information erasure, we consider a physical device that stores one bit of information, which is called a memory. Suppose that the memory is in contact with a single heat bath at temperature T $(= \beta^{-1})$. In the conventional setup, the memory is modeled by a particle in a symmetric double-well potential (see Fig. 4a), which has already been discussed in Sect. 1 (Fig. 1). The memory stores "0" ("1"), if the particle is in the left (right) well. The particle moves stochastically under the effect of a heat bath and can be described by, for example, an overdamped Langevin equation. Let x be the position of the particle, which is the physical degrees of freedom of this memory. We assume that the barrier between the wells is sufficiently high compared with the thermal energy T that thermal tunneling between the wells is negligible.

As a simpler model of the memory, the symmetric double-well potential can be replaced by two boxes with an equal volume (Fig. 4b), where the barrier of the double-well potential corresponds to the wall separating the boxes. In this two-box model, the memory stores "0" ("1"), if the particle is in the left (right) box.

In any setup (either the double-well model or the two-box model), the entire phase space, which we denote as X, represents the position of the particle. X is divided into two regions that represent computational states "0" and "1".

The information-erasure process with the two-box model is represented in Fig. 4c. We suppose that the memory is in local equilibrium in the individual boxes in the initial and final distributions. Since the two boxes have the same volume, the change in the Shannon entropy of the entire phase space by the information erasure is the same as that of the computational states:

$$\Delta S(X) = \Delta S(M) = -\ln 2. \tag{22}$$

From the second law (9) with $\Delta S(X) = -\ln 2$, we have

$$-Q \geq T \ln 2, \tag{23}$$

which implies that the heat emission $-Q$ from the memory is bounded from below by $T \ln 2$. This bound on heat emission is referred to as the Landauer bound, and inequality (23) is called the Landauer principle. Experimental verifications of the Landauer principle with a symmetric memory have been performed in, for example, Refs. [54, 59, 60, 62, 63].

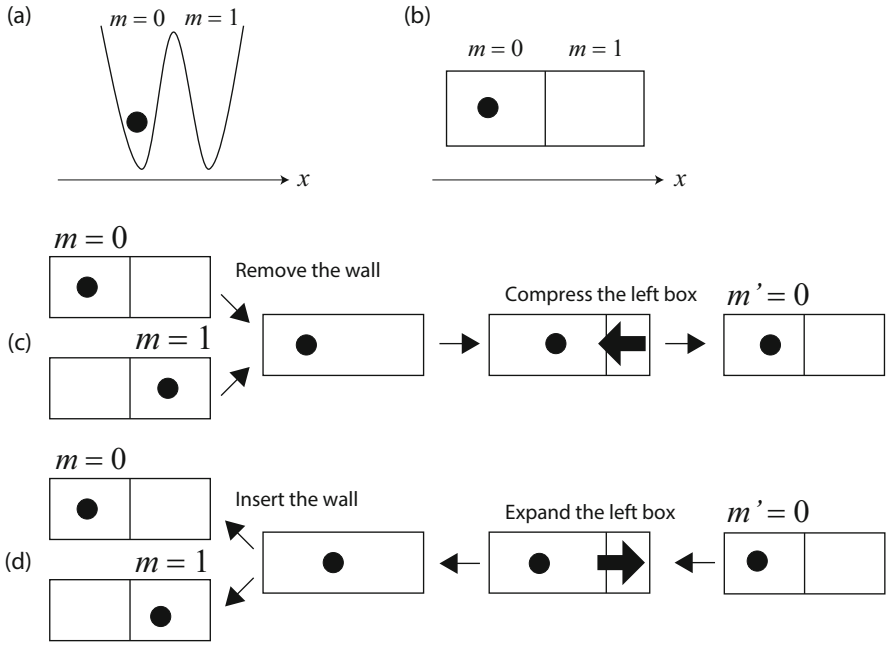

Fig. 4 Schematics of a symmetric memory. (**a**) The double-well potential model. The barrier at the center is assumed to be sufficiently high. (**b**) The two-box model. The left (right) box corresponds to the left (right) well of the double-well potential model. (**c**) Schematic of the information-erasure protocol with the two-box model. The initial computational state is $m = 0$ or $m = 1$ with probability $1/2$. The wall is instantaneously removed, which does not change the probability distribution of the particle. Then, the box is compressed to the left, and the final position of the wall is at the center. As a consequence, the final computational state is the standard state $m' = 0$ with unit probability. If the compression process is infinitely slow, this protocol is quasi-static and thermodynamically reversible, where the heat emission is given by $-Q = T \ln 2$. (**d**) The time-reversal of the quasi-static erasure protocol. The initial computational state is $m' = 0$, which is the final computational state of the erasure. The left box is first expanded infinitely slowly, and then the wall is inserted instantaneously. The final distribution is $m = 0$ or $m = 1$ with probability $1/2$, which is the initial distribution of the erasure. In this process, the heat absorption is given by $Q = T \ln 2$, which is equal and opposite to that in the erasure process

Let W be the work performed on the memory during the erasure. Since the internal energy does not change during the erasure, $W = -Q$ holds from the first law of thermodynamics. Therefore, the Landauer principle (23) can be rewritten as

$$W \geq T \ln 2, \tag{24}$$

which gives the fundamental lower bound of the work required for the information erasure in a symmetric memory.

The equality in (23) is achieved in the quasi-static limit, where the compression process in Fig. 4c is infinitely slow. In fact, such a quasi-static protocol is a special

Table 2 Summary of thermodynamic and logical reversibilities in the conventional setup of information erasure

	Quasi-static	Not quasi-static
Thermodynamically	Reversible	Irreversible
Logically	Irreversible	Irreversible
Heat emission $-Q$	$= T \ln 2$	$> T \ln 2$
Entropy production Σ	$= 0$	> 0

case of the reversible protocol in Fig. 2, and therefore the equality in (23) is achieved from the general argument in Sect. 3. In this case, the entropy production defined in (10) is zero: $\Sigma = 0$. We note that we can also directly compute that $-Q = T \ln 2$ in the infinitely slow limit, by using the equation of states of the single-particle gas, which will be discussed in Sect. 6 for more general situations (the present situation is $t = 1/2$ there).

Therefore, information erasure is thermodynamically reversible in the quasi-static limit. We note that the probability distribution is unchanged by the removal of the wall, which guarantees that the process is quasi-static. On the other hand, if information erasure is not quasi-static, the entropy production is positive: $\Sigma := \Delta S(X) - \beta Q > 0$. In this case, information erasure is thermodynamically irreversible.

To be more explicit, we show in Fig. 4d the time-reversal of the quasi-static information-erasure protocol with the two-box model, which indeed restores the probability distribution to the initial one, leading to the Shannon-entropy change $\Delta S(X)_{\rm reverse} = \ln 2$. In this time reversal process, the heat of $Q_{\rm reverse} = T \ln 2$ is absorbed from the heat bath during the expansion process, which has the inverse sign of the erasure process. We thus confirm that $\Sigma_{\rm reverse} := \Delta S(X)_{\rm reverse} - \beta Q_{\rm reverse} = 0$ in the time-reversal.

In short, logically irreversible information erasure can be performed in a thermodynamically reversible manner. Of course, this is totally consistent, given the different definitions of the two reversibilities. In fact, as also discussed in Sect. 1, logical reversibility cares only about the reversibility of the computational states, while thermodynamic reversibility is characterized by reversibility in the entire universe that consists of the memory and the heat bath.

From the entropic point of view, logical reversibility implies $\Delta S(M) = 0$, while thermodynamic reversibility implies $\Sigma := \Delta S(X) - \beta Q = 0$. These are definitely different, even when $\Delta S(M) = \Delta S(X)$ as in the present case.

In Table 2, we summarize the relationship between thermodynamic and logical reversibilities in the standard setup of information erasure.

6 Thermodynamics of Computation

In this section, we discuss a general framework of stochastic thermodynamics of computation. First, we remark that a physical state and a computational state are distinct concepts. In the standard setup of the Landauer principle in Sect. 5, the

physical state is the position of the particle, and thus is a continuous variable, while the computational state is "left" or "right" of the double well (representing "0" and "1"), and thus is a binary variable. In realistic situations of computation, a single computational state contains a huge number of microscopic physical states. This can be regarded as a coarse-graining of the physical phase space.

In general, we divide the physical phase space (i.e., the set of physical states) into several non-overlapping regions, where each region represents a computational state. Let X be the set of physical states, and M be the set of computational states, as in the previous sections. We consider a subset of X, written as X_m with index $m \in M$, which is the set of physical states that represent a computational state m. X is then divided as $X = \cup_m X_m$, where $X_m \cap X_{m'} = \phi$ for all $m \neq m'$ with the empty set ϕ.

We consider a probability distribution on the physical phase space. Let $P(x)$ be the probability of physical state $x \in X$, and $p(m)$ be that of computational state $m \in M$. Since all of $x \in X_m$ represent a single computational state m, we have

$$p(m) = \sum_{x \in X_m} P(x). \tag{25}$$

We then define the conditional probability of x under the condition that the computational state is m (i.e., $x \in X_m$):

$$P(x|m) = \begin{cases} P(x)/p(m) & (\text{if } x \in X_m), \\ 0 & (\text{otherwise}). \end{cases} \tag{26}$$

We next consider the Shannon entropy associated with this probability distribution. The Shannon entropy of the physical states is given by

$$S(X) := -\sum_x P(x) \ln P(x), \tag{27}$$

and the Shannon entropy of the computational states is given by

$$S(M) := -\sum_m p(m) \ln p(m). \tag{28}$$

We also consider the conditional entropy of X under the condition that the computational state is m:

$$S(X|m) := -\sum_{x \in X_m} P(x|m) \ln P(x|m), \tag{29}$$

which represents fluctuations of physical states inside a single computational state.

A crucial property of the non-overlapping division of the phase space is the corresponding decomposition of total (physical) entropy, which is represented as

$$S(X) = S(M) + S(X|M), \tag{30}$$

where

$$S(X|M) := \sum_m p(m) S(X|m) = -\sum_{m\in M}\sum_{x\in X_m} p(m)P(x|m)\ln P(x|m). \tag{31}$$

This decomposition is a general property of probability theory [77], and its proof is given by

$$\begin{aligned} S(X) &= -\sum_{x\in X} P(x)\ln P(x) \\ &= -\sum_{m\in M}\sum_{x\in X_m} p(m)P(x|m)\ln[p(m)P(x|m)] \\ &= -\sum_{m\in M} p(m)\ln p(m) - \sum_{m\in M}\sum_{x\in X_m} p(m)P(x|m)\ln P(x|m) \\ &= S(M) + S(X|M), \end{aligned} \tag{32}$$

where we used $\sum_{x\in X_m} P(x|m) = 1$ to obtain the third line. We note that $P(x|m)\ln P(x|m)$ does not diverge for all x and m.

Here, $S(X)$ represents the entire fluctuation in the physical phase space, which is related to the heat through the second law (9). On the other hand, $S(M)$ is the entropy of computational states, which is related to logical reversibility as discussed in Sect. 4. $S(X|M)$ represents the average of fluctuations inside the individual computational states. We refer to $S(X)$ as the physical entropy, $S(M)$ as the computational entropy, and $S(X|M)$ as the internal entropy.

We next consider dynamics on X that realizes a computation $\hat{C}$. The dynamics can be stochastic on the entire phase space X but should be deterministic on the computational space M in order to realize a deterministic computation. Such a situation is realistic in practical computations, because physical states thermally fluctuate inside individual computational states, even when the output of computation is deterministic.

We consider the change in the entropy during computation. Let m and m' be the initial and final computational states that are related deterministically as $m' = \hat{C}(m)$, and x and x' be the initial and final physical states that are related stochastically. We use notations X and X' (M and M') to refer to the probability variables of the initial and final physical (computational) states, respectively. The change in the Shannon entropies are then denoted as $\Delta S(X) := S(X') - S(X)$, $\Delta S(M) := S(M') - S(M)$, and $\Delta S(X|M) := S(X'|M') - S(X|M)$.

The second law (9) is represented by the total entropy as

$$\Delta S(X) \geq \beta Q, \tag{33}$$

which is equivalent to, via decomposition (30),

$$\Delta S(M) + \Delta S(X|M) \geq \beta Q. \tag{34}$$

Correspondingly, the entropy production (10) is decomposed as

$$\Sigma = \Delta S(M) + \Delta S(X|M) - \beta Q. \tag{35}$$

In the present setup, the nonequilibrium free energy is the same as Eq. (13):

$$F(X) := E(X) - TS(X) = E(X) - TS(M) - TS(X|M), \tag{36}$$

where $E(X)$ is the average energy of the memory. Then, the fundamental lower bound of the work W required for the computation is given by

$$W \geq \Delta F(X). \tag{37}$$

We consider the local canonical distribution inside a computational state m, which is given by

$$P_{\rm can}(x|m) := \begin{cases} e^{\beta(F_{\rm eq}(m) - E(x|m))} & (\text{if } x \in X_m), \\ 0 & (\text{otherwise}), \end{cases} \tag{38}$$

where $E(x|m)$ is the Hamiltonian for a given m, and

$$F_{\rm eq}(m) := -T \ln \sum_{x \in X_m} e^{-\beta E(x|m)} \tag{39}$$

is the local equilibrium free energy under the condition of m. If the memory is in local equilibrium inside individual computational states, the nonequilibrium free energy (36) reduces to [35]

$$F(X) = F_{\rm eq} - TS(M), \tag{40}$$

where $F_{\rm eq} := \sum_m p(m) F_{\rm eq}(m)$. If the initial and final distributions are local canonical, inequality (37) reduces to

$$W \geq \Delta F_{\rm eq} - T\Delta S(M). \tag{41}$$

In the rest of this section, we assume that the initial and final distributions are local canonical. In fact, this is a reasonable assumption, given that the time scale of global thermalization is much longer than that of local thermalization, because of the potential wall between the computational states.

◊◊◊

We now consider the role of the symmetry of the memory. We first consider the case that the memory is symmetric as in Sect. 5. In such a case, the local free energies of the two computational states are the same: $F_{\rm eq}(0) = F_{\rm eq}(1)$, and therefore $\Delta F_{\rm eq} = 0$ for any computation. Therefore, inequality (41) reduces to

$$W \geq -T\Delta S(M), \tag{42}$$

which is a general expression of the Landauer principle. In fact, the original Landauer principle (24) is a special case of inequality (42) with $\Delta S(M) = -\ln 2$. Inequality (42) has been directly verified in a recent experiment [62].

In terms of the internal entropy, the symmetry implies $S(X|0) = S(X|1)$ in local equilibrium, and therefore $\Delta S(X|M) = 0$. Then, inequality (34) reduces to

$$\Delta S(M) \geq \beta Q, \tag{43}$$

or equivalently,

$$-Q \geq -T\Delta S(M). \tag{44}$$

We next consider the case that the memory is asymmetric, where $\Delta F_{\rm eq} \neq 0$ and $\Delta S(X|M) \neq 0$ in general. If $\Delta S(X|M) \neq 0$, the entropy change in the computational states is not directly related to the heat (i.e., inequality (43) does not necessarily hold) [32, 34, 35]. In Fig. 5a, we show a simple example of a memory with an asymmetric double-well potential, where the left (right) well represents computational state "0" ("1").

As is the case for the symmetric memory, we can replace the double-well potential by two boxes (Fig. 5b). If the double-well potential is asymmetric, the volumes of the two boxes are not the same. Let t ($0 < t < 1$) be the ratio of the volume of the left box. If the memory is symmetric, $t = 1/2$. For $0 < t < 1/2$, the local free energies satisfy $F_{\rm eq}(0) > F_{\rm eq}(1)$, and the internal entropies satisfy $S(X|0) < S(X|1)$ in local equilibrium. We emphasize that the initial probability distribution of $m = 0$ and $m = 1$ is arbitrary (i.e., not necessarily $p(m = 0) = t$), because the memory can store any information.

We consider information erasure with the asymmetric memory. For simplicity, we assume that the initial distribution is $p(m = 0) = p(m = 1) = 1/2$. The Shannon-entropy change in the computational states by the information erasure is then given by $\Delta S(M) = -\ln 2$.

Figure 5c shows the optimal information-erasure protocol, which achieves the equality of the second law (34) [35]. A crucial point of this protocol is that the wall is first moved to the center infinitely slowly. Thanks to this process, the probability distribution of the particle (i.e., 1/2 for both left and right) does not change by the removal of the wall. (If we removed the wall without moving it to the center, the probability distribution would spontaneously relax towards the uniform distribution over the box, which makes the process thermodynamically irreversible and the entropy production positive.) This is in the same spirit as the protocol in Fig. 2.

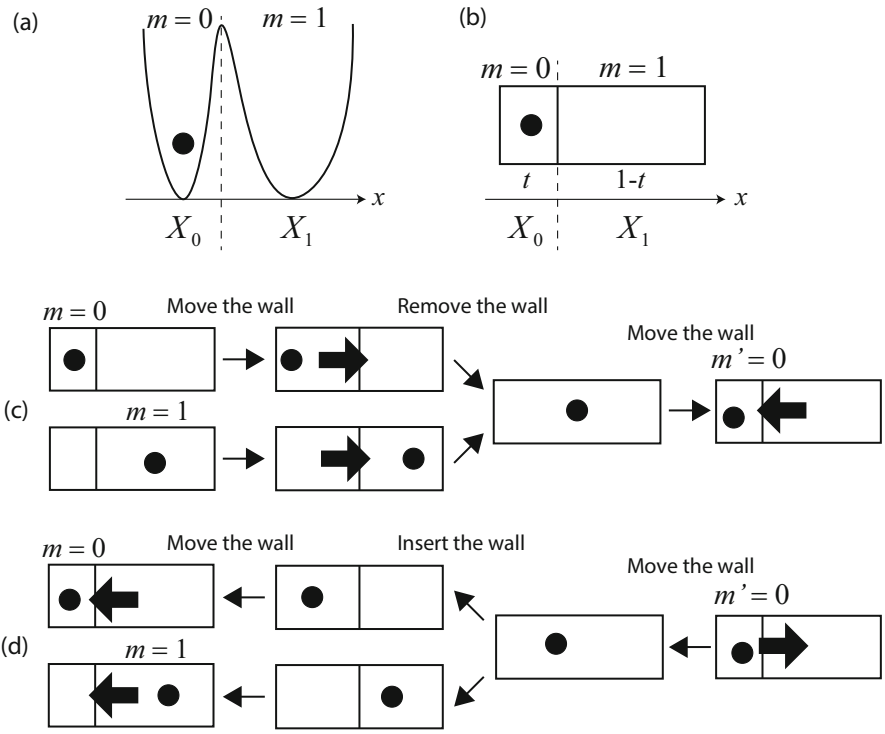

Fig. 5 Schematics of an asymmetric memory [35]. (**a**) An asymmetric double-well model. The phase-space volume of X_0 and X_1 is not equal. (**b**) The corresponding two-box model. The volumes of the left and right boxes are not equal. The volume ratio of the two boxes is given by $t : 1-t$ $(0 < t < 1)$. (**c**) The optimal information-erasure protocol with the asymmetric two-box model, which achieves the thermodynamic reversible condition $\Sigma = 0$. The initial computational state is $m = 0$ or $m = 1$ with probability $1/2$. The wall is moved to the center of the box infinitely slowly, and then removed instantaneously. The box is then compressed to the left infinitely slowly, so that the final volume ratio of the two boxes is the same as the initial one. The entire process of this erasure is quasi-static. (**d**) The time-reversal of the above quasi-static erasure protocol. The initial distribution of the time-reversal is the same as the final distribution of the erasure. The wall is first moved to the right most position infinitely slowly, and then a wall is instantaneously inserted at the center of the box. The inserted wall is then moved infinitely slowly, such that its final position is the same as its initial position of the erasure. In this process, the total heat absorption is given by $Q = T \ln 2 + (T/2) \ln(1 - t/t)$, which is equal and opposite to that in the erasure process

Then, the box is compressed from to the left infinitely slowly, and the final position of the wall returns to the initial one. The total entropy production is zero in this process, and thus it is thermodynamically reversible. To see the thermodynamic reversibility more explicitly, we illustrate the time-reversal of the above protocol in Fig. 5d.

We can also directly compute the heat emission for the protocol in Fig. 5c. We assume the equation of states of the single-particle gas, i.e., $PV = T$, where P is the

pressure and V is the volume of the box (and remind that $k_{\rm B} = 1$). The heat emission is then given by $-Q = W = -\int PdV$. We note that the work is not needed for the removal of the wall. Then, the heat emission during the entire process is given by

$$-Q = T \ln 2 + \frac{T}{2} \ln \frac{1-t}{t}. \tag{45}$$

On the other hand, we have $S(X|m=0) = S(X'|m'=0)$ and

$$S(X'|m'=0) - S(X|m=1) = \ln \frac{t}{1-t}, \tag{46}$$

and therefore,

$$\Delta S(X|M) := S(X'|m'=0) - \frac{1}{2}\left(S(X|m=0) + S(X|m=1)\right) = \frac{1}{2} \ln \frac{t}{1-t}. \tag{47}$$

Combining this with $\Delta S(M) = -\ln 2$, we obtain

$$\Delta S(M) + \Delta S(X|M) - \beta Q = -\ln 2 + \frac{1}{2} \ln \frac{t}{1-t} + \left(\ln 2 + \frac{1}{2} \ln \frac{1-t}{t} \right) = 0, \tag{48}$$

which achieves the equality in (34).

If $t = 1/2$, Eq. (45) reproduces that $-Q = T \ln 2$. On the other hand, if $t \neq 1/2$, we have $-Q \neq T \ln 2$. In particular, if $t > 1/2$, we have $-Q < T \ln 2$, which is below the Landauer bound (44). Of course, this does not contradict the second law. In such a case, the decrease in the computational entropy $\Delta S(M)$ is compensated for by the increase in the internal entropy $\Delta S(X|M)$. Information erasure with such an asymmetric memory has experimentally been demonstrated in Ref. [61].

In summary, heat emission is connected to the change in the total physical entropy of the memory (i.e., (i)+(ii) in Sect. 1), which is decomposed into the computational and internal entropies as in Eq. (30). If the change in the internal entropy is not zero, the computational entropy is not directly related to heat emission. This is the reason why the information erasure below the Landauer bound (44) is possible with an asymmetric memory, while the general bound (34) is always true.

7 Work Extraction and Reversibility with Feedback Control

We next consider work extraction from heat engines through feedback control by Maxwell's demon. As we will discuss below, the mutual information is the source of work extraction by the demon, and therefore we refer to such a heat engine as an information engine.

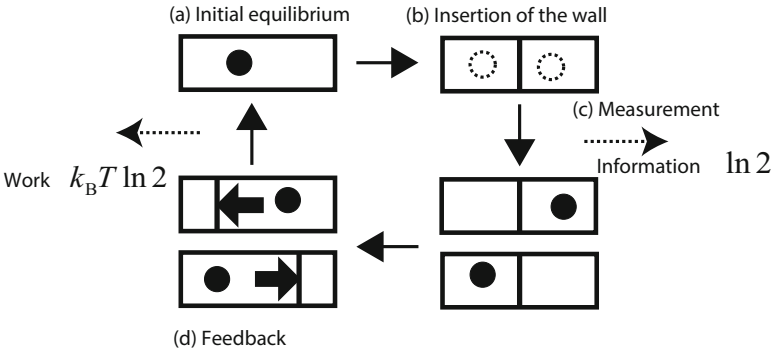

Fig. 6 Schematic of the Szilard engine. (**a**) The initial equilibrium distribution. (**b**) A wall is inserted at the center of the box. At this stage, we do not know in which side the particle is. (**c**) The demon measures the position of the particle (i.e., left or right). (**d**) If the particle of the engine is found in the left (right) box, the demon infinitely slowly expands the box to the right (left) so that the final distribution returns to the initial one. The work of $W_{\rm ext} = T \ln 2$ is extracted from this expansion. Since the direction of the expansion depends on the measurement outcome (left or right), this process is regarded as feedback control by the demon

In this section, we do not explicitly formulate the memory of the demon itself. Instead, we only regard the demon as an external agent that affects the probability distribution of the engine through the measurement. The full analysis of the total system of engine and demon is postponed to the next section.

We first consider the Szilard engine, which is a simple model of an information engine (see Fig. 6). The Szilard engine consists of a Brownian particle (or a molecule) in a box that is attached to a single heat bath at temperature $T = \beta^{-1}$. During the process depicted in Fig. 6, the demon obtains one bit ($= \ln 2$) of information corresponding to left or right, and performs feedback control. Then, the work of $W_{\rm ext} = T \ln 2 > 0$ is extracted from the engine, and the same amount of heat $Q = T \ln 2$ is absorbed from the bath. The amount of the work is calculated by the same manner as in Sect. 6. Since the dynamics of the engine is cyclic, this positive work extraction apparently violates the second law. However, if we take into account the memory of the demon, then the total system is not cyclic, and therefore this is not a violation of the second law. As will be discussed in Sect. 8, we can understand more quantitatively the consistency between the demon and the second law by taking into account the mutual information between the engine and the demon.

We now consider a general upper bound of the extractable work with feedback control. We assume that the engine is in contact with a single heat bath at temperature $T = \beta^{-1}$. Let $P(x)$ be the probability distribution of the engine immediately before the measurement by the demon. We note that we use notation x (and X) to describe the engine, instead of the memory of the demon; we use notation y for the measurement outcome obtained by the demon.

We suppose that the error of the measurement is described by the conditional probability $P(y|x)$, which is the probability of obtaining outcome y under the condition that the true state of the engine is x. If the measurement is error-free, we have $P(y|x) = \delta_{xy}$ with δ_{xy} being the Kronecker's delta (or the delta function if the state variable is continuous). The joint probability of x and y is given by $P(x, y) = P(y|x)P(x)$, and the unconditional probability of y is given by $P(y) = \sum_x P(x, y)$. From the Bayes rule, the conditional probability of x under the condition of outcome y is given by

$$P(x|y) = \frac{P(x, y)}{P(y)}. \tag{49}$$

Correspondingly, the conditional entropy of X under the condition of a particular y is given by

$$S(X|y) := -\sum_x P(x|y) \ln P(x|y). \tag{50}$$

Its ensemble average over all y is

$$S(X|Y) := \sum_y P(y) S(X|y) = -\sum_{xy} P(x, y) \ln P(x|y) = S(XY) - S(Y), \tag{51}$$

where $S(XY) := -\sum_{xy} P(x, y) \ln P(x, y)$ is the Shannon information of the joint distribution.

After the measurement, the protocol to control the engine depends on y, which is the characteristic of feedback control. By noting that the initial distribution of the engine is given by $P(x|y)$ under the condition of outcome y, the second law of stochastic thermodynamics (9) can apply to the conditional distribution:

$$S(X'|y) - S(X|y) \geq \beta Q_y, \tag{52}$$

where Q_y is the heat absorption with y, and $S(X'|y)$ is the conditional entropy in the final distribution of the engine. By taking the ensemble average over all y, we have

$$S(X'|Y) - S(X|Y) \geq \beta Q, \tag{53}$$

where $Q := \sum_y P(y) Q_y$.

Before proceeding further, we here discuss mutual information, which quantifies a correlation between two probability variables. The mutual information between X and Y is defined as

$$I(X : Y) := S(X) + S(Y) - S(XY) = \sum_{x,y} P(x, y) \ln \frac{P(x, y)}{P(x)P(y)}. \tag{54}$$

It immediately follows that

$$I(X:Y) = S(X) - S(X|Y) = S(Y) - S(Y|X). \tag{55}$$

The mutual information satisfies the following inequalities:

$$0 \leq I(X:Y) \leq \min\{S(X), S(Y)\}, \tag{56}$$

where $I(X:Y) = 0$ holds if and only if the two systems are not correlated (i.e., $P(x, y) = P(x)P(y)$).

Going back to the second law (53), it is rewritten as, by using Eq. (55),

$$\Delta S(X) - \beta Q \geq -\Delta I, \tag{57}$$

where $\Delta S(X) := S(X') - S(X)$ and $\Delta I := I(X':Y) - I(X:Y)$. We note that if the feedback control works, $I(X':Y) < I(X:Y)$ should hold (and thus $\Delta I < 0$). In fact, in the case of the Szilard engine, $I(X:Y) = \ln 2$ and $I(X':Y) = 0$ hold, because there is no remaining correlation after the entire process. In general, since the correlation is also decreased by dissipation to the environment, $-\Delta I$ gives an upper bound of the information that is utilized by feedback control. By noting that $\Delta S(X) - \beta Q$ is nonnegative in the absence of feedback control, inequality (57) implies that we can reduce the entropy of the system by using feedback control, where the mutual information is the resource of the entropy reduction.

We consider the nonequilibrium free energy of X, defined in the same manner as Eq. (13):

$$F(X) := E(X) - TS(X), \tag{58}$$

where $E(X)$ is the average energy of the engine. We then rewrite inequality (57) as

$$W \geq \Delta F(X) + T\Delta I. \tag{59}$$

By defining the extracted work $W_{\mathrm{ext}} := -W$, we have

$$W_{\mathrm{ext}} \leq -\Delta F(X) - T\Delta I. \tag{60}$$

The right-hand side above can be further bounded as

$$W_{\mathrm{ext}} \leq -\Delta F(X) + TI(X:Y), \tag{61}$$

where we used $I(X':Y) \geq 0$. Inequality (61) implies that additional work can be extracted up to the mutual information obtained by the measurement.

We consider a special case that the initial distribution of the engine is canonical and the final distribution is also canonical under the condition of y. More precisely, the final Hamiltonian can depend on y, which we denote by $E(x'|y)$, and the

final distribution is given by $P(x'|y) = e^{\beta(F_{\rm eq}(X'|y)-E(x'|y))}$, where $F_{\rm eq}(X'|y) := -T\ln\sum_{x'} e^{-\beta E(x'|y)}$ is the final equilibrium free energy with y. Let $F_{\rm eq}(X)$ be the initial equilibrium free energy as usual. We then have

$$S(X) = \beta(E(X) - F_{\rm eq}(X)),\ \ S(X'|y) = \beta(E(X'|y) - F_{\rm eq}(X'|y)), \tag{62}$$

where $E(X'|y) := \sum_{x'} P(x'|y)E(x'|y)$ that gives $E(X') = \sum_y P(y)E(X'|y)$. We define the change in the equilibrium free energy by

$$\Delta F_{\rm eq}(X) := \sum_y P(y)F_{\rm eq}(X'|y) - F_{\rm eq}(X). \tag{63}$$

By substituting Eq. (62) into inequality (59), we have

$$W \geq \Delta F_{\rm eq}(X) - TI(X:Y), \tag{64}$$

or equivalently,

$$W_{\rm ext} \leq -\Delta F_{\rm eq}(X) + TI(X:Y). \tag{65}$$

We emphasize that inequality (64) or (65) is exactly equivalent to (59) under the assumption that the initial and final distributions are (conditional) canonical, where we did not drop $TI(X':Y)$. In fact, to obtain inequality (64) or (65) from (59), $TI(X':Y)$ is just absorbed into the definition of $\Delta F_{\rm eq}(X)$ in Eq. (63). On the other hand, we dropped $TI(X':Y)$ to obtain (61) from (59).

In the case of the Szilard engine, we have $W_{\rm ext} = T\ln 2$, $\Delta F_{\rm eq}(X) = 0$, and $I(X:Y) = \ln 2$. Therefore, the equality in (65) is achieved in the Szilard engine.

We note that inequality (65) has been derived in Refs. [41, 43]. The role of mutual information in thermodynamics has been experimentally demonstrated in, for example, Refs. [56, 67].

◊◊◊

We consider thermodynamic reversibility with feedback control [81, 82]. We remember that the second law with feedback control is given by inequality (53), which is the ensemble average of inequality (52). Here, inequality (52) is equivalent to the second law (9) under the condition of y. Therefore, it is reasonable to adopt the following definition [82]:

Definition (Thermodynamic Reversibility with Feedback) In the presence of feedback control, thermodynamic reversibility of the engine is achieved if and only if the equality in (53) is achieved, or equivalently, the equality in (52) is achieved for all y.

In the rest of this section, we work on this definition of thermodynamic reversibility. We note, however, that this definition does not concern the reversibility

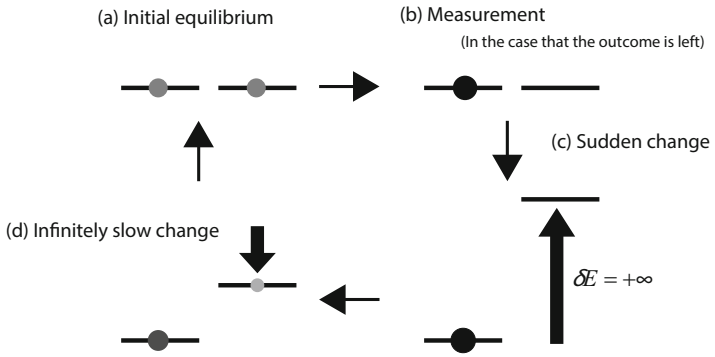

Fig. 7 An analogue of the Szilard engine. (**a**) In the initial equilibrium distribution, the particle is in the left or the right site with probability 1/2. (**b**) The demon performs the measurement of the position of the particle, and obtains ln 2 of information. The case that the outcome is "left" is shown in this figure. (**c**) If the particle is found in the left (right) site, the demon suddenly changes the energy level of the right (left) site to $+\infty$. This is analogous to the insertion of the wall of the original Szilard engine. In this sudden change, we do not need any work. (**d**) The demon infinitely slowly lowers the energy level of the right (left) site to the original level, from which $T \ln 2$ of the work is extracted

of the memory of the demon during the measurement process. In fact, in this section we have just formulated the measurement process as the modification of the probability distribution from $P(x)$ to $P(x|y)$, without explicitly considering dynamics of the memory. The full treatment of the memory will be discussed in Sect. 8 in detail.

By remembering the argument in Sect. 3, thermodynamic reversibility is achieved by the protocol in Fig. 2, where we now replace the distribution $P(x)$ by the conditional one $P(x|y)$, and also the potential $V(x)$ by the y-dependent one $V(x|y)$. In other words, thermodynamic reversibility with feedback control is achieved if we adjust the potential $V(x|y)$ such that the conditional distribution $P(x|y)$ becomes always the canonical distribution of $V(x|y)$. In particular, we need to switch the potential immediately after the measurement, because the distribution is suddenly changed from $P(x)$ to $P(x|y)$ by the measurement. We again remark that this consideration neglects reversibility of the measurement process itself.

We revisit the Szilard engine as a special example. Since the Szilard engine achieves the equality in (65) as mentioned above, the Szilard engine is thermodynamically reversible. We can directly confirm that the Szilard engine is always in the canonical distribution under a particular measurement outcome.

To see this point clearer, let us consider a simple analogue of the Szilard engine, illustrated in Fig. 7. In this model, the particle is in one of the two sites with the same energy, which is in contact with a single heat bath at temperature T $(= \beta^{-1})$. The information gain ln 2 and the work extraction $T \ln 2$ in this model are the same as those in the Szilard engine, implying the thermodynamic reversibility of this model. It is obvious that this model is always in the conditional canonical distribution during the entire process.

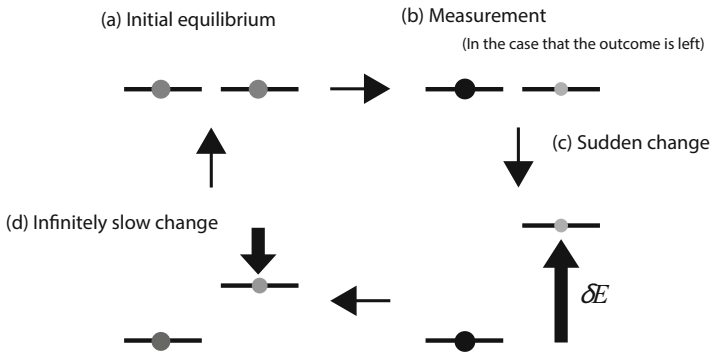

Fig. 8 The Szilard-type engine with measurement error [82]. (**a**) In the initial equilibrium distribution, the particle is in the left or the right site with probability 1/2. (**b**) The demon performs the measurement of the position of the particle, and obtains the mutual information (66). The case that the outcome is "left" is shown in this figure. (**c**) If the particle is found in the left (right) site, the demon suddenly changes the energy level of the right (left) site such that the energy difference is given by δE. For this sudden change, a positive amount of work is performed if $\varepsilon \neq 0$. (**d**) The demon infinitely slowly lowers the energy level of the right (left) site to the original level, from which a positive amount of work is extracted

We can generalize this model by incorporating a measurement error [82], which is illustrated in Fig. 8. We suppose that the error rate of the measurement is given by ε ($0 \leq \varepsilon \leq 1$); the conditional probabilities are given by $P(y|x) = 1 - \varepsilon$ ($x = y$) and $P(y|x) = \varepsilon$ ($x \neq y$) with x, y being "right" or "left." In this case, the mutual information obtained by this measurement is

$$I(X : Y) = \ln 2 + \varepsilon \ln \varepsilon + (1 - \varepsilon) \ln(1 - \varepsilon). \tag{66}$$

Immediately after the measurement, we have $P(x|y) = 1 - \varepsilon$ ($x = y$) and $P(x|y) = \varepsilon$ ($x \neq y$). To achieve thermodynamic reversibility, we need to make $P(x|y)$ the canonical distribution for all y. Consider the case that $y =$"left" as illustrated in Fig. 8. (The same argument applies to the case that $y =$"right".) The demon switches the energy level of the right site to make the energy difference $\delta E = -T \ln(\varepsilon/(1 - \varepsilon))$ so that $P(x|y)$ becomes canonical (Fig. 8c):

$$\frac{e^{-\beta\delta E}}{1 + e^{-\beta\delta E}} = \varepsilon, \quad \frac{1}{1 + e^{-\beta\delta E}} = 1 - \varepsilon. \tag{67}$$

The work extraction by this switching is given by $-\varepsilon\delta E$ on average, because the particle is pushed up if it is in the right site.

The demon next lowers the energy level of the right site infinitely slowly, and the final distribution is the same as the initial one (Fig. 8d). The extracted work during this process is given by $T \ln(2/(1 + e^{-\beta\delta E}))$, because the extracted work equals the minus of the equilibrium free-energy change in this situation (i.e., the free energy

after the sudden switching is $-T\ln(1+e^{-\beta\delta E})$ and that in the final distribution is $-T\ln 2$).

The total work extracted from the entire process is then given by

$$W_{\rm ext} = -\varepsilon\delta E + T\ln\frac{1}{1+e^{-\beta\delta E}} = T\left(\ln 2 + \varepsilon\ln\varepsilon + (1-\varepsilon)\ln(1-\varepsilon)\right). \quad (68)$$

We note that $\Delta F_{\rm eq}(X) = 0$ in the entire process. Therefore, the equality in (65) is achieved, i.e., $W_{\rm ext} = TI(X:Y)$, and thus we confirm that this protocol achieves the thermodynamic reversibility.

This type of the Szilard engine with measurement error has been proposed in Ref. [82], and experimentally demonstrated in Ref. [56] by using a single electron box. Other models that achieve the thermodynamic reversibility with feedback have been discussed in Refs. [46, 83–86].

8 Entropy Balance in Maxwell's Demon

In the previous section, we did not explicitly consider the measurement as a physical process, but as the modification of the probability distribution from $P(x)$ to $P(x|y)$. In particular, we did not consider the entropy production in the memory of the demon itself.

In this section, we explicitly consider stochastic thermodynamics of the entire system of the engine X and the memory of the demon Y [50, 51]. Specifically, we focus on the entropy balance during the measurement and the feedback processes by explicitly considering the memory as a physical system. In this respect, we will reproduce the second law (53) with feedback control from a slightly different viewpoint from Sect. 7. We also discuss the fundamental energy cost required for the measurement process.

As a preliminary, we consider general dynamics of the bipartite system X and Y in the presence of a heat bath at temperature $T = \beta^{-1}$. The entropy production in the total system is given by

$$\Sigma(XY) := \Delta S(XY) - \beta Q_{XY}, \quad (69)$$

where $\Delta S(XY)$ is the change in the joint Shannon entropy, and Q_{XY} is the heat absorbed by the total system. We can also define the entropy production in the subsystem X by

$$\Sigma(X) := \Delta S(X) - \beta Q_X, \quad (70)$$

where $\Delta S(X)$ is the change in the Shannon entropy of X, and Q_X is the heat absorbed by X from the heat bath. In the same manner, we define

$$\Sigma(Y) := \Delta S(Y) - \beta Q_Y. \tag{71}$$

In many physical situations (e.g., a bipartite Markov jump process and a Langevin system with two variables driven by independent noise), we can suppose that the heat is additive:

$$Q_{XY} = Q_X + Q_Y. \tag{72}$$

On the other hand, the Shannon entropy is generally not additive, and the mutual information appears:

$$\Delta S(XY) = \Delta S(X) + \Delta S(Y) - \Delta I(X:Y). \tag{73}$$

By using Eqs. (72) and (73), the total entropy production is decomposed as

$$\Sigma(XY) = \Sigma(X) + \Sigma(Y) - \Delta I(X:Y), \tag{74}$$

where the total entropy production is not additive too, because of the mutual information term. This observation is crucial to understand the consistency between Maxwell's demon and the second law, as discussed below. We emphasize that the second law of thermodynamics always applies to the total entropy production:

$$\Sigma(XY) \geq 0. \tag{75}$$

Correspondingly, a process is thermodynamically reversible if and only if $\Sigma(XY) = 0$.

We note that the terminology of the entropy "production" for the subsystems (i.e., $\Sigma(X)$ and $\Sigma(Y)$) is a little bit of an abuse. More precisely, $\Sigma(X)$ is the sum of the entropy increase in X and that in the bath associated with dynamics of X. Strictly speaking, the terminology of "production" should be reserved for the entropy increase of the total system, not for that of a subsystem. In the following, however, for the sake of simplicity, we refer to $\Sigma(X)$ and $\Sigma(Y)$ just as the entropy production of the subsystems.

◊◊◊

We now consider stochastic thermodynamics of the measurement and feedback processes (see Fig. 9 for a schematic). We suppose that subsystem X is an engine measured and controlled by the demon, and subsystem Y plays the role of the memory of the demon. The Szilard engine discussed above is a special case of this setup; Fig. 10 shows the dynamics of the Szilard engine along with the memory of the demon. To avoid too much complication, we do not explicitly formulate the computational states of the demon in the following, while it is straightforward to consider them [37].

We first consider the measurement process. Before the measurement, the system and the demon are not correlated, and the mutual information is zero. Let x be the initial state of the engine and y_0 the initial state of the demon. During the

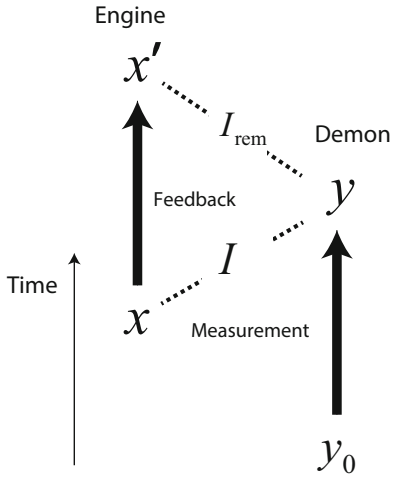

Fig. 9 Schematic of the measurement and the feedback processes, where x, x' (y_0, y) represent the initial and the final states of the engine (the memory of the demon). The initial correlation between the engine and the memory is assumed to be zero. After the measurement of the engine by the demon, a correlation is established, which is represented by the mutual information I. Feedback control is performed by using the measurement outcome y, and the remaining correlation after feedback is I_{rem}

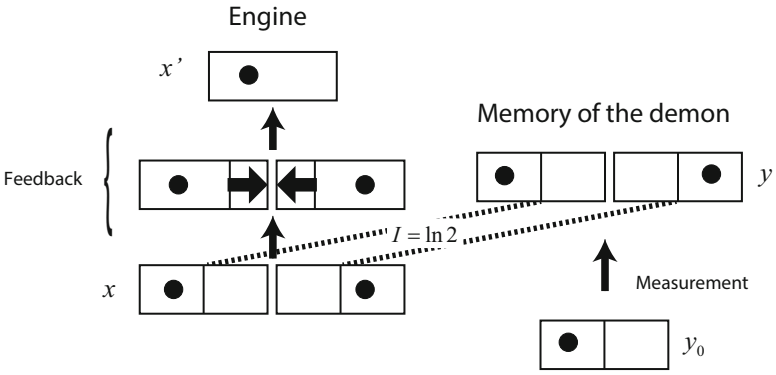

Fig. 10 A schematic of the Szilard engine and the memory of the demon, which is a special case of Fig. 9. Here, both of the engine and the memory are represented by the two boxes with a particle

measurement, the dynamics of the demon depends on the initial state x of the system. For simplicity, we assume that the measurement reads out the instantaneous value of x, and the system does not evolve during the measurement. After the measurement, the state of the demon, denoted as y, is correlated with x. Here, y is supposed to be equivalent to the measurement outcome in Sect. 7.

Let I be the mutual information between x and y. The mutual-information change during the measurement is given by

$$\Delta I_{\rm meas} = I, \tag{76}$$

which is positive if the demon gains information. From Eq. (74), the entropy production in the total system during the measurement is given by

$$\Sigma(XY)_{\rm meas} = \Sigma(X)_{\rm meas} + \Sigma(Y)_{\rm meas} - I. \tag{77}$$

From the assumption that the system does not evolve during the measurement, $\Sigma(X)_{\rm meas} = 0$. Therefore, we obtain

$$\Sigma(XY)_{\rm meas} = \Sigma(Y)_{\rm meas} - I. \tag{78}$$

Since the second law applies to the total entropy production, $\Sigma(XY)_{\rm meas} \geq 0$, we obtain

$$\Sigma(Y)_{\rm meas} \geq I. \tag{79}$$

This implies that the entropy production of the memory during the measurement is bounded from below by the mutual information.

In terms of the nonequilibrium free energy (13), we rewrite inequality (79) as

$$W_{\rm meas} \geq \Delta F(Y)_{\rm meas} + TI, \tag{80}$$

where $\Delta F(Y)_{\rm meas} := \Delta E(Y)_{\rm meas} - T\Delta S(Y)_{\rm meas}$. Inequality (80) reveals the fundamental lower bound of the energy cost for the measurement. Here, TI on the right-hand side comes from the right-hand side of Eq. (79), and represents the additional energy cost to obtain the mutual information I. This inequality has been derived in Refs. [35, 50].

We next consider the feedback process, where the dynamics of the engine depends on the measurement outcome y. For simplicity, we assume that the memory does not evolve during the measurement (i.e., y remains unchanged). After the feedback, the final state of the system is x', and the remaining correlation between x' and y is denoted as $I_{\rm rem}$. The mutual-information change during feedback is then given by

$$\Delta I_{\rm fb} = I_{\rm rem} - I. \tag{81}$$

This is negative if the obtained information is used during the feedback by the demon as discussed in Sect. 7. We note that I, $I_{\rm rem}$, and $\Delta I_{\rm fb}$ respectively equal $I(X:Y)$, $I(X':Y)$, and ΔI in the notations of Sect. 7.

From Eq. (74), the entropy production in the total system during the feedback is given by

$$\Sigma(XY)_{\rm fb} = \Sigma(X)_{\rm fb} + \Sigma(Y)_{\rm fb} + I - I_{\rm rem}. \tag{82}$$

From the assumption that the memory does not evolve during the feedback, $\Sigma(Y)_{\rm fb} = 0$. Therefore, we obtain

$$\Sigma(XY)_{\rm fb} = \Sigma(X)_{\rm fb} + I - I_{\rm rem}. \tag{83}$$

Again since the second law applies to the total entropy production, $\Sigma(XY)_{\rm fb} \geq 0$, we obtain

$$\Sigma(X)_{\rm fb} \geq -(I - I_{\rm rem}). \tag{84}$$

This implies that the entropy production of the system during the feedback can be negative up to the minus of the used information by the feedback. We note that inequality (84) is equivalent to inequality (57) in Sect. 7, where $\Sigma(X)_{\rm fb}$ equals $\Delta S(X) - \beta Q$ in the notation of Sect. 7. In the case of the Szilard engine, $\Sigma(X)_{\rm fb} = -\ln 2$. Such reduction of the entropy is the bare essential of the role of Maxwell's demon.

We note that thermodynamic reversibility is achieved if and only if, for the measurement and the feedback processes,

$$\Sigma(XY)_{\rm meas} = 0, \quad \Sigma(XY)_{\rm fb} = 0, \tag{85}$$

respectively. A model of Maxwell's demon that satisfies both of these reversibility conditions has been proposed in Ref. [85].

As a side remark, we consider information erasure from the memory after the feedback process. In the erasure process, the memory does not interact with the engine, but solely goes back to the initial distribution only in contact with the heat bath. In this process, the second law is given by

$$\Sigma(Y)_{\rm erase} := \Delta S(Y)_{\rm erase} - \beta Q_{\rm erase} \geq 0, \tag{86}$$

which is nothing but the (generalized) Landauer principle (37). In the quasi-static limit, we have $\Sigma(Y)_{\rm erase} = 0$. In terms of the work and the nonequilibrium free energy, inequality (86) is rewritten as

$$W_{\rm erase} \geq \Delta F(Y)_{\rm erase}. \tag{87}$$

We assume the complete information erasure, in which after the information erasure, the probability distribution and the Hamiltonian of the memory completely return to the initial ones before the measurement. This assumption is satisfied if the memory is in the standard computational state with local equilibrium, before the measurement and after the erasure. In this case, $\Delta F(Y)_{\rm meas} = -\Delta F(Y)_{\rm erase}$. Therefore, by summing up inequalities (80) and (87), we obtain [35]

$$W_{\rm meas} + W_{\rm erase} \geq TI. \tag{88}$$

Inequality (88) is the trade-off relation between the work for the measurement and that for the erasure, and sets the fundamental lower bound of the energy cost required for the memory. We remark that the lower bound of (88) is given only by the mutual information, but does not depend on the details of the memory (e.g., symmetric or asymmetric). This mutual-information term exactly compensates for the additionally extractable work by feedback control (i.e., the mutual-information term in inequality (65)).

◊◊◊

We now summarize the key observation in the foregoing argument. First of all, the measurement and feedback processes are *individually* consistent with the second law, because $\Sigma(XY) \geq 0$ holds for the individual processes. In this respect, there is not any contradiction between the second law and Maxwell's demon.

The apparent "paradox" of Maxwell's demon would stem from the negative entropy production of the engine, $\Sigma(X)_{\rm fb} < 0$. However, the second law must apply to the total system, and therefore the negative entropy production of the subsystem is not a contradiction. If we take into account the change in the mutual information, by adding it to $\Sigma(X)_{\rm fb}$ as $\Sigma(X)_{\rm fb} + (I - I_{\rm rem})$, we recover the total entropy production $\Sigma(XY)_{\rm fb}$ that is always nonnegative.

In the case of the Szilard engine, $\Sigma(X)_{\rm fb} = -\ln 2$ and $I - I_{\rm rem} = \ln 2$. Therefore, the total entropy production is just zero: $\Sigma(XY)_{\rm fb} = -\ln 2 + \ln 2 = 0$, which implies that the Szilard engine is a reversible information engine. Table 3 summarizes the entropy balance of the Szilard engine for the case that the measurement, the feedback, and the erasure processes are all quasi-static.

As discussed above, an information-erasure process can follow the feedback process. We emphasize that, however, we do not necessarily need to consider information erasure to understand the consistency between the demon and the second law.

Table 3 The entropy balance of the Szilard engine, where X is the engine and Y is the demon

	$\Sigma(XY)$	$\Sigma(X)$	$\Sigma(Y)$	ΔI
Measurement	0	0	$\ln 2$	$\ln 2$
Feedback	0	$-\ln 2$	0	$-\ln 2$
Erasure	0	0	0	0

Here, we assumed that all the processes (i.e., measurement, feedback, and erasure) are quasi-static

9 Concluding Remarks

In this article, we have only focused on the second law of thermodynamics at the level of the ensemble average, with which we have clarified the concept of reversibilities and the entropy production. However, stochastic thermodynamics has much richer aspects, which we did not discuss so far. In the following, we will briefly summarize some important topics beyond the scope of this article.

Fluctuation Theorem One of the most important discoveries in stochastic thermodynamics is the fluctuation theorem [11, 12, 14–18]. Roughly speaking, we consider the stochastic version of the entropy production σ, which gives $\Sigma = \langle \sigma \rangle$ with $\langle \cdots \rangle$ being the ensemble average. Then, the fluctuation theorem (or more precisely, the integral fluctuation theorem or the Jarzynski equality) is given by

$$\langle e^{-\sigma} \rangle = 1, \tag{89}$$

which implies that the second law of thermodynamics can be represented by an equality, if we take into account fluctuations of the entropy production. By using the convexity of the exponential function, we have $\langle e^{-\sigma} \rangle \geq e^{-\langle \sigma \rangle}$. Therefore, Eq. (89) reproduces the usual second law $\langle \sigma \rangle \geq 0$. We note that the fluctuation-dissipation theorem and its generalization to nonlinear responses can be obtained from the fluctuation theorem (89) [87, 88].

Thermodynamics of information can be formulated at the level of the stochastic entropy production, and thus the fluctuation theorem can be generalized by incorporating the mutual information [43, 50, 51].

Autonomous Demons In Sects. 7 and 8, we have discussed Maxwell's demon that performs a single measurement-feedback process. We can extend the second law and the fluctuation theorem to multiple measurement-feedback processes [45, 46], and further to situations that measurement and feedback are performed autonomously and continuously in time [89–101]. Here, the informational quantities that characterize continuous information flow, such as the transfer entropy [102] and the learning rate (or just the "information flow") [95, 101], play crucial roles.

There is also another formulation of autonomous demons based on the concept of information reservoirs [103–108]. These two approaches (the autonomous measurement-feedback approach and the information reservoir approach) are shown equivalent in general [109]. It has also been shown that there is an exact mapping between these approaches for a typical model [110], based on the concept of partial entropy production [97]. We note that other informational quantities, such as the Kolmogorov-Sinai entropy, have been investigated in the context of thermodynamics [111, 112].

Application to Biological Systems Interesting applications of thermodynamics of information, especially the theory of autonomous demons, are also found in biophysics. In fact, living cells perform autonomous information processing based

on biochemical reactions; Thermodynamics of information in biochemical systems is now an active emerging field [113–124].

Quantum Thermodynamics and Quantum Information We have focused on classical thermodynamics and classical information so far, while stochastic thermodynamics also applies to quantum systems [109, 125–130]. Quantum analogues of the Szilard engine have been proposed [131, 132], and the role of quantum information in thermodynamics has been intensively investigated [35, 41, 44, 133–140]. Furthermore, several experiments on thermodynamics of information have been performed in the quantum regime [67–69]. We also note that there is another interesting approach to quantum thermodynamics, called thermodynamic resource theory [141, 142].

Ultimate Origin of the Information-Thermodynamics Link Last but not least, the fundamental origin of the information-thermodynamics link is yet to be fully understood based on quantum mechanics. Throughout this manuscript, we have assumed that there exists a large heat bath in thermal equilibrium, specifically in the canonical distribution. However, the microscopic characterization of thermal equilibrium is quite nontrivial, because a typical pure quantum state [143] and even a single energy eigenstate [144] can behave as thermal. In this context, the eigenstate-thermalization hypothesis (ETH) has been considered to be a plausible mechanism of thermalization in isolated quantum systems [144]. Based on the ETH, the second law and the fluctuation theorem have been proved in the short time regime for isolated quantum many-body systems where the heat bath is initially in a single energy eigenstate [145].

◊◊◊

In these decades, there has been significant progress in stochastic thermodynamics, which has led to the modern theory of thermodynamics of information. Stochastic thermodynamics is still quite a hot field, and thermodynamics of information would further lead to the fundamental understanding of the interplay between physics and information.

Acknowledgements The author is grateful to Masahito Ueda, John Bechhoefer, Jordan M. Horowitz, and Naoto Shiraishi for a lot of valuable comments, and to Christian Van den Broeck, Massimiliano Esposito, and Udo Seifert for valuable suggestions. This work is supported by JSPS KAKENHI Grant No. JP16H02211 and No. JP25103003.

References

1. J.C. Maxwell, *Theory of Heat* (Appleton, London, 1871)
2. L. Szilard, Z. Phys. **53**, 840 (1929)
3. C. Shannon, Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
4. H.S. Leff, A.F. Rex (eds.), *Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing* (Princeton University Press, New Jersey, 2003)
5. L. Brillouin, J. Appl. Phys. **22**, 334 (1951)

6. R. Landauer, IBM J. Res. Dev. **5**, 183 (1961)
7. C.H. Bennett, Int. J. Theor. Phys. **21**, 905 (1982)
8. W.H. Zurek, Nature **341**, 119 (1989)
9. R. Landauer, Science **272**, 1914 (1996)
10. H. Matsueda, E. Goto, K.-F. Loe, RIMS Kôkyûroku **1013**, 187 (1997)
11. D.J. Evans, E.G.D. Cohen, G.P. Morris, Phys. Rev. Lett. **71**, 2401 (1993)
12. G. Gallavotti, E.G.D. Cohen, Phys. Rev. Lett. **74**, 2694 (1995)
13. B. Gaveau, L. Schulman, Phys. Lett. A **229**, 347–353 (1997)
14. C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997)
15. G.E. Crooks, Phys. Rev. E **60**, 2721 (1999)
16. C. Jarzynski, J. Stat. Phys. **98**, 77 (2000)
17. D.J. Evans, D.J. Searles, Adv. Phys. **51**, 1529 (2002)
18. U. Seifert, Phys. Rev. Lett. **95**, 040602 (2005)
19. R. Kawai, J.M.R. Parrondo, C. Van den Broeck, Phys. Rev. Lett. **98**, 080602 (2007)
20. M. Esposito, C. Van den Broeck, Europhys. Lett. **95**, 40004 (2011)
21. K. Sekimoto, *Stochastic Energitics* (Springer, Berlin/Heidelber, 2010)
22. C. Jarzynski, Ann. Rev. Condens. Matter Phys. **2**, 329 (2011)
23. U. Seifert, Rep. Prog. Phys. **75**, 126001 (2012).
24. K. Kawasaki, J.D. Gunton, Phys. Rev. A **8**, 2048 (1973)
25. J. Schnakenberg, Rev. Mod. Phys. **48**, 571 (1976)
26. G.N. Bochkov, Yu. E. Kuzovlev, Zh. Eksp. Teor. Fiz. **72**, 238–247 (1977)
27. C. Van den Broeck, Selforganization by nonlinear irreversible processes, in *Proceedings of the Third International Conference*, Kühlungsborn, GDR, 18–22 March 1985, ed. by W. Ebeling, H. Ulbricht (Springer, Berlin, 1985), pp. 57–61
28. C.Y. Mou, J.-L. Luo, G. Nicolis, J. Chem. Phys. **84**, 7011 (1986)
29. J.M.R. Parrondo, J.M. Horowitz, T. Sagawa, Nat. Phys. **11**, 131 (2015)
30. K. Shizume, Phys. Rev. E **52**, 3495 (1995)
31. B. Piechocinska, Phys. Rev. A **61**, 062314 (2000)
32. M.M. Barkeshli, arXiv:cond-mat/0504323 (2005)
33. R. Dillenschneider, E. Lutz, Phys. Rev. Lett. **102**, 210601 (2009)
34. S. Turgut, Phys. Rev. E **79**, 041102 (2009)
35. T. Sagawa, M. Ueda, Phys. Rev. Lett. **102**, 250602 (2009); **106**, 189901(E) (2011)
36. D. Reeb, M.M. Wolf, New J. Phys. **16**, 103011 (2014)
37. T. Sagawa, J. Stat. Mech. P03025 (2014)
38. H. Touchette, S. Lloyd, Phys. Rev. Lett. **84**, 1156 (2000)
39. H. Touchette, S. Lloyd, Phys. A **331**, 140 (2004)
40. K.H. Kim, H. Qian, Phys. Rev. E **75**, 022102 (2007)
41. T. Sagawa, M. Ueda, Phys. Rev. Lett. **100**, 080403 (2008)
42. F.J. Cao, M. Feito, Phys. Rev. E **79**, 041118 (2009)
43. T. Sagawa, M. Ueda, Phys. Rev. Lett. **104**, 090602 (2010)
44. T. Sagawa, Prog. Theor. Phys. **127**, 1 (2012)
45. J.M. Horowitz, S. Vaikuntanathan, Phys. Rev. E **82**, 061120 (2010)
46. T. Sagawa, M. Ueda, Phys. Rev. E **85**, 021104 (2012)
47. D. Abreu, U. Seifert, Phys. Rev. Lett. **108**, 030601 (2012)
48. S. Lahiri, S. Rana, A.M. Jayannavar, J. Phys. A: Math. Theor. **45**, 065002 (2012)
49. S. Still, D.A. Sivak, A.J. Bell, G.E. Crooks, Phys. Rev. Lett. **109**, 120604 (2012)
50. T. Sagawa, M. Ueda, Phys. Rev. Lett. **109**, 180602 (2012)
51. T. Sagawa, M. Ueda, New J. Phys. **15**, 125012 (2013)
52. S. Ciliberto, Phys. Rev. X **7**, 021051 (2017)
53. S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, M. Sano, Nat. Phys. **6**, 988 (2010)
54. A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, E. Lutz, Nature **483**, 187–189 (2012)
55. E. Roldan, I.A. Martinez, J.M.R. Parrondo, D. Petrov, Nat. Phys. **10**, 457 (2014)
56. J.V. Koski, V.F. Maisi, T. Sagawa, J.P. Pekola, Phys. Rev. Lett. **113**, 030601 (2014)

57. J.V. Koski, A. Kutvonen, I.M. Khaymovich, T. Ala-Nissila, J.P. Pekola, Phys. Rev. Lett. **115**, 260602 (2015)
58. K. Chida, S. Desai, K. Nishiguchi, A. Fujiwara, Nat. Commun. **8**, 15310 (2017)
59. A. Bérut, A. Petrosyan, S. Ciliberto, Europhys. Lett. **103**, 60002 (2013)
60. Y. Jun, M. Gavrilov, J. Bechhoefer, Phys. Rev. Lett. **113**, 190601 (2014)
61. M. Gavrilov, J. Bechhoefer, Phys. Rev. Lett. **117**, 200601 (2016)
62. M. Gavrilov, R. Chétrite, J. Bechhoefer, Proc. Natl. Acad. Sci. USA **114**, 11097–11102 (2017)
63. J. Hong, B. Lambson, S. Dhuey, J. Bokor, Sci. Adv. **11**, e1501492 (2016)
64. M. Lopez-Suarez, I. Neri, L. Gammaitoni, Nat. Commun. **7**, 12068 (2016)
65. J.P.P. Silva et al., Proc. R. Soc. A **472**, 20150813 (2016)
66. M.D. Vidrighin, O. Dahlsten, M. Barbieri, M.S. Kim, V. Vedral, I.A. Walmsley, Phys. Rev. Lett. **116**, 050401 (2016)
67. P.A. Camati et al., Phys. Rev. Lett. **117**, 240502 (2016)
68. N. Cottet et al., Proc. Natl. Acad. Sci. USA **114**, 7561–7564 (2017)
69. Y. Masuyama, K. Funo, Y. Murashita, A. Noguchi, S. Kono, Y. Tabuchi, R. Yamazaki, M. Ueda, Y. Nakamura, arXiv:1709.00548 (2017)
70. H.B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, 2nd edn. (Wiley, New York, 1985)
71. E.H. Lieb, J. Yngvasonb, Phys. Rep. **314**, 669 (1999)
72. D. Kondepudi, I. Prigogine, *From Heat Engines to Dissipative Structures* (Wiley, New York, 1998)
73. M. Gavrilov, J. Bechhoefer, Europhys. Lett. **114**, 50002 (2016)
74. N. Shiraishi, K. Saito, H. Tasaki, Phys. Rev. Lett. **117**, 190601 (2016)
75. F. Curzon, B. Ahlborn, Am. J. Phys. **43**, 22 (1975)
76. M. Esposito, R. Kawai, K. Lindenberg, C. Van den Broeck, Phys. Rev. Lett. **105**, 150603 (2010)
77. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991)
78. N.G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd edn. (North-Holland Personal Library, 2007)
79. W.C. Gardiner, *Handbook of Stochastic Methods*, 3rd edn. (Springer, Berlin, 2004)
80. C. Moore, S. Mertens, *The Nature of Computation* (Oxford University Press, Oxford, 2011)
81. K. Jacobs, Phys. Rev. A **80**, 012322 (2009)
82. J.M. Horowitz, J.M.R. Parrondo, Europhys. Lett. **95**, 10005 (2011)
83. J.M. Horowitz, J.M.R. Parrondo, New J. Phys. **13**, 123019 (2011)
84. D. Abreu, U. Seifert, Europhys. Lett. **94**, 10001 (2011)
85. J.M. Horowitz, T. Sagawa, J.M.R. Parrondo, Phys. Rev. Lett. **111**, 010602 (2013)
86. C. Kwon, Phys. Rev. E **95**, 042103 (2017)
87. D. Andrieux, P. Gaspard, J. Stat. Mech. Theor. Exp. P02006 (2007)
88. K. Saito, Y. Utsumi, Phys. Rev. B **78**, 115429 (2008)
89. A.E. Allahverdyan, D. Janzing, G. Mahler, J. Stat. Mech. Theor. Exp. P09011 (2009)
90. Y. Fujitani, H. Suzuki, J. Phys. Soc. Jpn. **79**, 104003 (2010)
91. P. Strasberg, G. Schaller, T. Brandes, M. Esposito, Phys. Rev. Lett. **110**, 040601 (2013)
92. S. Ito, T. Sagawa, Phys. Rev. Lett. **111**, 180603 (2013)
93. D. Hartich, A.C. Barato, U. Seifert, J. Stat. Mech. P02016 (2014)
94. T. Munakata, M.L. Rosinberg, Phys. Rev. Lett. **112**, 180601 (2014)
95. J.M. Horowitz, M. Esposito, Phys. Rev. X **4**, 031015 (2014)
96. J.M. Horowitz, H. Sandberg, New J. Phys. **16**, 125007 (2014)
97. N. Shiraishi, T. Sagawa, Phys. Rev, E **91**, 012130 (2015)
98. N. Shiraishi, S. Ito, K. Kawaguchi, T. Sagawa, New J. Phys. **17**, 045012 (2015)
99. S. Yamamoto, S. Ito, N. Shiraishi, T. Sagawa, Phys. Rev. E **94**, 052121 (2016)
100. M.L. Rosinberg, J.M. Horowitz, Eur. Phys. Lett. **116**, 10007 (2016)
101. D. Hartich, A.C. Barato, U. Seifert, Phys. Rev. E **93**, 022116 (2016)
102. T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000)
103. D. Mandal, C. Jarzynski, Proc. Natl. Acad. Sci. USA **109**, 11641 (2012)

104. S. Deffner, C. Jarzynski, Phys. Rev. X **3**, 041003 (2013)
105. A.C. Barato, U. Seifert, Phys. Rev. Lett. **112**, 090601 (2014)
106. A.C. Barato, U. Seifert, Phys. Rev. E **90**, 042150 (2014)
107. N. Merhav, J. Stat. Mech. P06037 (2015)
108. A.B. Boyd, D. Mandal, J.P. Crutchfield, New J. Phys. **18**, 023049 (2016)
109. P. Strasberg, G. Schaller, T. Brandes, M. Esposito, Phys. Rev. X **7**, 021003 (2017)
110. N. Shiraishi, T. Matsumoto, T. Sagawa, New J. Phys. **18**, 013044 (2016)
111. A.B. Boyd, J.P. Crutchfield, Phys. Rev. Lett. **116**, 190601 (2016)
112. A.B. Boyd, D. Mandal, P.M. Riechers, J.P. Crutchfield, Phys. Rev. Lett. **118**, 220602 (2017)
113. G. Lan, P. Satori, S. Neumann, V. Sourjik, Y. Tu, Nat. Phys. **8**, 422–428 (2012)
114. P. Mehta, D.J. Schwab, Proc. Natl. Acad. Sci. USA **109**, 17978 (2012)
115. A.C. Barato, D. Hartich, U. Seifert, Phys. Rev. E **87**, 042104 (2013)
116. A.C. Barato, D. Hartich, U. Seifert, New J. Phys. **16**, 103024 (2014)
117. A.H. Lang, C.K. Fisher, T. Mora, P. Mehta, Phys. Rev. Lett. **113**, 148103 (2014)
118. P. Sartori, L. Granger, C.F. Lee, J.M. Horowitz, PLoS Compt. Biol. **10**, e1003974 (2014)
119. S. Ito, T. Sagawa, Nat. Commun. **6**, 7498 (2015)
120. T. J. Kobayashi, Y. Sughiyama, Phys. Rev. Lett. **115**, 238102 (2015)
121. P. Sartori, S. Pigolotti, Phys. Rev. X **5**, 041039 (2015)
122. T.E. Ouldridge, P.R. ten Wolde, Phys, Rev. Lett. **118**, 158103 (2017)
123. T.E. Ouldridge, C.C. Govern, P.R. ten Wolde, Phys, Rev. X **7**, 021004 (2017)
124. T. Matsumoto, T. Sagawa, arXiv:1711.00264 (2017)
125. J. Kurchan, arXiv:cond-mat/0007360 (2000)
126. H. Tasaki, arXiv:cond-mat/0009244 (2000)
127. M. Esposito, U. Harbola, S. Mukamel, Rev. Mod. Phys. **81**, 1665 (2009)
128. T. Campisi, P. Hanggi, P. Talkner, Rev. Mod. Phys. **83**, 771 (2011)
129. T. Sagawa, arXiv:1202.0983 (2012); Chapter of *Lectures on Quantum Computing, Thermodynamics and Statistical Physics*. Kinki University Series on Quantum Computing (World Scientific, Singapore, 2012)
130. À. M. Alhambra, L. Masanes, J. Oppenheim, C. Perry, Phys. Rev. X **6**, 041017 (2016)
131. W.H. Zurek, in *Frontiers of Nonequilibrium Statistical Physics*, ed. by G.T. Moore, M.O. Scully. NATO ASI Series (Series B: Physics), vol. 135 (Springer, Boston, 1986). arXiv:quant-ph/0301076
132. S.W. Kim, T. Sagawa, S. De Liberato, M. Ueda, Phys. Rev. Lett. **106**, 070401 (2011)
133. S. Lloyd, Phys. Rev. A **39**, 5378 (1989)
134. M.A. Nielsen, C.M. Caves, B. Schumacher, H. Barnum, Proc. R. Soc. Lond. A **454**, 277 (1998)
135. Y. Morikuni, H. Tasaki, J. Stat. Phys. **143**, 1 (2011)
136. L. del Rio, J. Aberg, R. Renner, O. Dahlsten, V. Vedral, Nature **474**, 61 (2011)
137. K. Funo, Y. Watanabe, M. Ueda, Phys. Rev. E **88**, 052121 (2013)
138. H. Tajima, Phys. Rev. E **88**, 042143 (2013)
139. J.J. Park, K.-H. Kim, T. Sagawa, S.W. Kim, Phys. Rev. Lett. **111**, 230402 (2013)
140. J. Goold, M. Huber, A. Riera, L. del Rio, P. Skrzypzyk, J. Phys. A: Math. Theor. **49**, 143001 (2016)
141. M. Horodecki, J. Oppenheim, Nat. Commun. **4**, 2059 (2013)
142. F.G.S.L. Brandão, M. Horodecki, N.H.Y. Ng, J. Oppenheim, S. Wehner, Proc. Natl. Acad. Sci. **112**, 3275 (2015)
143. S. Popescu, A.J. Short, A. Winter, Nat. Phys. **2**, 754–758 (2006)
144. M. Rigol, V. Dunjko, M. Olshanii, Nature **452**, 854–858 (2008)
145. E. Iyoda, K. Kaneko, T. Sagawa, Phys. Rev. Lett. **119**, 100601 (2017)

The Thermodynamics of Computation: A Contradiction

Wolfgang Porod

Contents

1 Introduction

It is an intriguing notion that the laws of physics might imply limits on computation. Any computation in the real world inevitably involves some physical structure and processes that realize the symbolic computational steps. In particular, the question of fundamental limits of the energy requirement for computation has attracted significant attention in the literature [1–3]. In an influential paper, Landauer in 1961 used thermodynamic arguments to conclude that logically irreversible operations, such as erasure, necessarily lead to the generation of heat in the computing process, thus being a fundamental source of energy dissipation. This paper [4] has spawned many subsequent studies and created a school of thought that intimately links thermodynamics and computation.

However, there also have been dissenting voices, including this author [5, 6], who have not only questioned the use of thermodynamic arguments for computation, but

W. Porod (✉)
Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA
e-mail: porod@nd.edu

C. S. Lent et al. (eds.), *Energy Limits in Computation*,
https://doi.org/10.1007/978-3-319-93458-7_4

argued that thermodynamics does not even apply to computation. There also have been other dissenting voices, pointing out other issues, including Refs. [7–13]. It is the purpose of this chapter to focus on the connection (or lack thereof) between thermodynamics and computation.

Boltzmann's concept of entropy and its implication for reversibility have been deeply controversial from the very beginning and remain so until today. It is perhaps no surprise then that invoking entropy in the context of energy dissipation in computation has stirred controversy. This controversy also is evident by the different positions presented in the various chapters in this book, and we will leave it to the reader to come to their own conclusion. Rather than presenting a resolution of these conflicting views, the goal of this book has been to get to the core of these controversial arguments.

A central concept in these arguments is the concept of entropy. Unfortunately, the vast majority of the literature on this topic simply invokes "entropy," without distinguishing between different definitions of entropy, such as physical entropy or information entropy. This lack of clarity regarding what is actually meant by entropy has significantly contributed to the confusion in the literature on this topic. We will carefully examine the meaning and limits of validity of these two forms of entropy, and we will show that information entropy is relevant for computation, and physical (thermodynamic) entropy is not. As a consequence, thermodynamics does not apply to computation, and the thermodynamics of computation is a contradiction in terms. Thus, the title for this chapter.

Another concept that is frequently invoked without clear definition is the concept of reversibility. There is physical reversibility, which is related to the time reversibility of a physical process, and there is logical reversibility, which is related to the invertibility of a logic operation. The concept of reversibility also is used in thermodynamics and the "arrow of time" (second law of thermodynamics). Moreover, reversibility also is invoked in the context of adiabatic processes. Clearly, arguments about reversibility in one context should not be used to reach conclusions about reversibility in another context, and we will be careful not to so in this chapter.

This chapter is structured in the following way: In order to discuss the energy requirements of computation in a meaningful way, one first needs to define what one means by "computation." Here, the model of computation will be that of a Turing machine, since this is what Landauer and Bennett had used. We will then present Landauer's influential arguments, which use thermodynamics to discuss the energetics of bit operations, using the concept of entropy. We will point out that there are different forms of entropy, and we will discuss in detail the distinction between physical (thermodynamic) entropy and information entropy. This will lead us to the conclusion that information entropy applies to computation, but physical entropy does not.

2 Model of Computation

In order to discuss the physics of a computational process in a meaningful way, one first has to define what one means by computation. In this section, we will do just that and define what we mean by computation, and we will also say what we do not mean by computation in this context.

Landauer's and Bennett's model of computation is that of a Turing machine, which is an abstract computing machinery introduced by Alan Turing in the late 1940s [14]. Specifically, a Turing machine contains a data-storage device, the so-called tape, which is composed of a sequence of segments, each of which contains a symbol (bit) value. These symbol (bit) values can be changed by a bit-manipulating device, the so-called head, which can assume certain distinct states according to a set of rules. The computation proceeds in discrete steps, which involve reading and writing bit values, and changing the state and the location of the head.

An elementary operation of a Turing machine has the following form: In the beginning, the head will be in a certain initial state, denoted by H_i, and it will be located above some segment along the tape. The head will then read the symbol (bit) value of the tape at that location, which we will denote by S_i. Once the head has obtained the information about the current bit value, it will execute a rule, which changes the symbol value to its final state, denoted by S_f, and the head itself changes its state to H_f. Finally, the head will move either to the right or to the left (or not at all), and the next computation step can begin with the new symbol and head states.

Formally, an elementary operation of a Turing machine maps an initial head/symbol state to a final symbol/head state, with the additional head movement, M:

$$(H_i, S_i) \rightarrow \left(S_f, H_f, M\right)$$

We note that this is a 2-to-3 mapping, which is logically irreversible.

Alternatively, elementary Turing-machine operations can be written as quintuples of the form:

$$\left(H_i, S_i; S_f, H_f, M\right)$$

A set of such mappings, or quintuples, defines the set of rules for a Turing-machine program, and it has been shown that Turing machines are computationally universal.

While the Turing machine is an abstract computational model, it is also useful for considering sources of energy requirements for computation. Symbol (bit) values have to be maintained in the presence of noise, otherwise the data will be randomized. What is also needed is some physical mechanism, representing the head, that is capable of reading, manipulating, and writing bit values, and of changing its own state. All of this will require more or less energy, depending upon the specific physical system used to realize a symbolic bit.

Of special importance here is the read operation by which the head acquires the current bit value, and which then determines what operation the head will execute next. In other words, a Turing machine consists of elementary steps, where each step is determined by the current state of the computation (data). In order for the computation to stay on the correct path, these read operations are crucial, or the Turing machine will perform a random walk in the computational space.

It has been well recognized that these read operations, which the head needs in order to "know" what to do next, are similar to the read operations Maxwell's demon needs in order to "know" when to open his trap door. We will return to a discussion of this connection to Maxell's demon later in this chapter.

After having discussed the Turing-machine model of computation, which will be the basis for the arguments in this chapter, we now briefly present a couple of other computational models, which have nothing to do with the thermodynamics of computation. Unfortunately, the literature also contains papers that use these models of computation in this context, further confusing the issues of interest here.

One such model is the billiard-ball model of computation [15]. The presence, or absence, of balls is interpreted as bit information, and bit operations can be accomplished by having balls bounce off each other, or physical structures, such as mirrors or walls. Note that these physical structures have to be in place before the balls start, and someone had to design the layout these mirrors and walls for a specific path of the balls. For some logic gates, additional balls are required to set the functionality of such a gate. Note that these extra balls have to be aimed and timed perfectly, in order to arrive precisely when they are needed to bounce off the "computing" balls. In other words, the state of the computation has to be known along all the ball trajectories before the "computation" even starts. Since the state of the computation has to be known for all times, this includes the end of computation. In other words, the outcome of the "computation" is known before the "computation" even starts. As such, this model of computation is very different from the Turing machine. This billiard-ball model is more like an automaton that is designed for a particular task, and it can only execute that special task. On the other hand, a Turing machine is general purpose, and does not have to be specifically designed for a particular computation.

Since the trajectories of ideal balls are reversible (physically and thus logically), this model of computation is reversible (both physically and logically). Also, if the motion of the balls is assumed to be frictionless, and the collisions are assumed to be without deformation or transfer of energy, this machinery operates without dissipation (albeit by assumption).

There exists a large literature on adiabatic reversible computing, and a few references are given here [16–19]. In particular, adiabatic charging techniques can significantly reduce power consumption using mainstream conventional CMOS technology. The main idea here is to design circuits such that no large currents are allowed to flow across resistors while charging (or discharging) capacitors. This work certainly is valid as it directly addresses important source of power dissipation, i.e. Joule heating in resistors, and tries to minimize these sources of wasting energy. However, we argue, this work has nothing to do with Landauer's

arguments involving logical reversibility and thermodynamics. In both instances the general notion of reversibility appears, albeit with very different meaning. An adiabatic charging process uses clocking structures, which make it physically reversible, and since information is encoded in the state of the capacitor, this makes it also logically reversible. In other words, logical reversibility is a consequence of physical reversibility. On the other hand, Landauer's arguments claim that physical reversibility is a consequence of logical reversibility, which is very different.

3 Landauer and the School of Dissipationless Computation

In this section, we will state Landauer's original arguments, which led him to conclude that logically irreversible processes require dissipation because of thermodynamics.

In his influential 1961 paper, Landauer argued that logical irreversibility is the source of dissipation. Erasure re-sets a bit to a certain (the erased) state (say, the 0 state) [4]. Since the bit could have been initially in either the 0 or 1 state, he argued that erasure is a 2-to-1 mapping. He then used thermodynamic arguments to conclude that such a contraction of state (phase) space necessarily requires expenditure of energy. Specifically, he argued that the entropy of the initial state is $S_i = k_B \ln 2$, and the entropy of the final state is $S_f = k_B \ln 1 = 0$. According to thermodynamics then, this change in bit entropy requires an expenditure of energy equal to $E = k_B T \ln 2$.

Bennett then showed in 1973 that computation can be done in a logically reversible fashion by using extra bits [20]. Logically irreversible operations, i.e. many-to-few mappings, can be made logically reversible by the addition of extra bits to have same-to-same mappings. He argued that these extra bits have to be erased at the end of the computation, and that this is the fundamental source of dissipation (according to Landauer's original argument). In order to avoid the erasure of such random bits, he devised a three-step computational process to accomplish this:

- In Step 1, the computation is performed, and one ends up with the result of the computation, plus the extra bits from the intermediate results that have to be erased.
- In Step 2, the results of the computation are copied to a separate tape, which he claims can be done in a dissipationless fashion.
- In Step 3, the computation from Step 1 is run backwards, and in the end, one ends up with the initial state of the computation, thus effectively erasing the random bits by un-computing them.

These arguments, which link logical irreversibility to energy dissipation through thermodynamics, have found a wide following, which we shall call the *School of Dissipationless Computation*. An early review of this school of thought was given

by Bennett in his paper entitled "The Thermodynamics of Computation—a Review" [21], as well as in subsequent reviews [22, 23]. A more recent review is given by Lutz and Ciliberto in their 2015 article in Physics Today [24].

There also have been critical voices, including this author, who have questioned Landauer's use of thermodynamic arguments for the energetics of bit operations [5]. In fact, we will show that thermodynamics does not apply to bit operations, which calls into question the very foundation of the *School of Dissipationless Computation*. This is the main message of this chapter.

4 The School of Dissipationless Computation Contains Inconsistencies and Contradictions

Before going into a detailed criticism of the thermodynamic arguments invoked by *The School of Dissipationless Computation* (which is what we will do below), we first want to point out that their arguments lead to inconsistent and unphysical conclusion, and even contain contradictions.

A basic tenet of this school of thought is that "erasure" is special (fundamentally requiring dissipation), and distinct from other bit operations. As already stated above, Landauer argues that during the process of erasure, an initial bit value, which could have been "zero" or "one," is transformed into one specific bit state, say "zero," which is interpreted as the erased bit state. However, any bit operation is like this. An initial bit state, which could have been in one state or the other, is transformed into the final bit state. All a bit can do is to exist in one of two states, which applies to both the initial and final state of a bit operation. At the physical bit level, there is nothing special about "erasure"; it is a bit operation like any other. The only difference is at the human level, i.e. how a particular bit operation is interpreted. Clearly, simply calling a bit operation "erasure" does not change the underlying physics. More specifically, simply calling some bit operations "erasure" and others not will not make some dissipative and others not.

This school of thought wants to make us believe that computing (forward) is dissipative, whereas computing forward and then backward is not (Bennett's three tape construction). Please note that our arguments here, as were Bennett's, are for the Turing-machine model of computation, and not for energy-recycling adiabatic processes. As already stated above, the basic rationale for Bennett's three stage construction is Landauer's notion that "erasure" requires dissipation, whereas "computing" does not. In order for this construction to make sense, the "computing" steps have to be truly dissipationless, or one would dissipate twice as much computing forward and then backward, as opposed to just computing forward. It shall be emphasized here that such dissipationless computing is a mere assertion, without any physical basis. Clearly, any meaningful discussion of dissipationless computing would require a discussion of the physics of these bit operations themselves, but there is no such discussion in this school of thought.

The assertion of dissipationless "computing" steps has further consequences. Since each computing step in a Turing machine requires a read operation followed by a bit operation, both have to be dissipationless. But, if read operations do not require any energy, and bit operations do not require any energy, one can transform a bit in an arbitrary initial state into a specific final state without requiring any energy. Specifically, one can write a program for a Turing machine which reads an initial bit value, dissipationless by assumption, and which then simply writes a chosen final state (say, "zero," regardless of the initial state), again dissipationless by assumption. In this fashion, one can have a program for a Turing machine that "computes" an arbitrary sequence of bit values into a sequence of just zero's. In other words, if one believes the school's assertion of dissipationless computing, one can "compute" the "erasure" of a random tape in a dissipationless fashion, in clear contradiction to their own assertion that "erasure" fundamentally requires dissipation.

5 Physical Entropy Versus Information Entropy, and Other Forms of Entropy

We will now discuss various forms of entropy, especially physical (thermodynamic) entropy and information entropy, and we will show that they have very different meanings. Central to a discussion of entropy are the concepts of degrees of freedom and of phase space.

Thermodynamics was developed in the 1870s by Ludwig Boltzmann to describe the experimentally observable macroscopic behavior of gases as a consequence of the underlying random thermal motion of the gas particles, which is not accessible to experiment. Even though the details of that thermal motion are unknown, it still can be described as the microscopic dynamics of the gas molecules, which can be idealized as classical point particles for this purpose. In this classical-mechanics picture, the instantaneous particle locations and velocities represent the microscopic degrees of freedom. For a gas with N molecules, the set of locations and velocities for each of these N molecules defines one particular microstate of the gas. The set of all possible microscopic configurations that are compatible with the macroscopic boundary conditions defines the phase space of the gas.

A central quantity in thermodynamics is the entropy, which represents an average over all microscopic configurations that correspond to a macrostate. Specifically, the thermodynamic entropy, S, is defined as:

$$S = -k_B \, \Sigma \, p_i \, \ln \, p_i$$

Here, k_B is the Boltzmann constant, and p_i is the probability with which a particular microstate, labeled by the subscript "i," contributes to the macrostate. The sum is over all microstates "i" in the phase space that corresponds to the macrostate.

For the special case that all microstates contribute with equal probability, the entropy is given by:

$$S = k_B \ \ln \ W$$

Here, W is the number of microstates that correspond to the gas' macrostate, and $p_i = 1/W$. By the way, this famous formula is inscribed on Boltzmann's gravestone in the Zentralfriedhof in Vienna.

The entropy, which is a measure of the thermal motion of the gas molecules, is directly related to heat energy by:

$$\Delta E = T \Delta S$$

T is the absolute temperature. The above relation states that a change in entropy, ΔS, leads to a change in heat energy, ΔE, which can be used to perform work, such as in a steam engine. This is a statement of the conservation of energy, where heat energy can be transformed into other (more useful) forms of energy. This is the essence of the first law of thermodynamics.

It shall be pointed out that the sum over microstates, which was defined mathematically above, also has a very physical meaning. This sum essentially represents an average over all microstates during an experiment, which yields the experimentally observed macrovariables. Strictly (mathematically) speaking, this sum is over all possible microstates in the whole phase space. Clearly, in any experimental situation, it will not be possible for the gas molecules to actually find themselves in all these possible configurations. However, the thermal motion is so rapid, and there are so many collisions, that for practical purposes the actual microstates provide a representative sample of all possible microstates. This means that one does not have to know what particular microstates contributed to the averages observable in a particular experiment, and one can simply perform thermodynamic averages over all possible states. There are two important points here: One, in order for the sum over microstates to have physical meaning, these microstates have to actually, physically contribute to the average (and not just mathematically). Second, for practical purposes, the particular subset of microstates during an actual experimental situation provides a representative sample of all possible microstates.

We shall also point out that it is possible to construct experimental conditions where the microstates that lead to the observed macroscopic quantities are not a representative sample of the thermodynamic phase space. For example, if experiments are performed at sufficiently short times scales, which do not allow proper averaging, non-thermodynamic behavior can be observed.

We now turn our attention to information entropy. In analogy to a gas container with N molecules, one can define analogous quantities for a sequence of N bits, such as the tape of a Turing machine with length N. Now, the "microscopic" degrees of freedom are the bit values, and a "microstate" is a particular sequence of these N bit values. The combination of all possible 2^N bit sequences defines the "phase space."

In complete analogy to thermodynamics, one can now define an entropy-like function by performing weighted sums over all possible bit sequences (microstates). We shall call this entropy-like function information entropy. Furthermore, and again in analogy to thermodynamics, one can define an energy-like function, that relates changes in information entropy to this such-defined "energy."

It is rather obvious that information entropy has a very different physical meaning than thermodynamic entropy. There is no random thermal dynamics for sequences of bits, as there is for gas molecules. Averages over microstates in a gas have a very clear physical meaning, as they directly determine the value of macrovariables. Averages over bit sequences certainly can be performed mathematically, but they do not have the same physical meaning as thermodynamic averages. We will discuss the connection between energy and these two forms of entropy in the following section.

We would like to point out that the degrees of freedom discussed so far for a bit sequence, i.e. the bit values, are the information-bearing degrees of freedom that carry meaning for the state of the computation (Turing machine). For any physical implementation of bits, say the presence or absence of charge on a capacitor, there also are additional degrees of freedom, such as the detailed location of these charges, or how these charges couple to their microscopic environment by phonons, etc. We shall call these additional degrees of freedom, which are part of the thermal environment, non-information-bearing degrees of freedom. For the bit information, we only care about the presence of charge, and not about the microscopic details of how these charges couple to the thermal background. By their very definition, one does not have control over these non-information-bearing degrees of freedom, and all one can do is form thermodynamic averages. In other words, thermodynamics applies to the non-information-bearing degrees of freedom, but not to the information-bearing degrees of freedom.

We now conclude this section by making an analogy, which might appear silly, but which serves to capture the essence of the important distinction between information entropy and physical entropy.

One might liken bit configurations to arrangements of objects, such as chairs in a conference room, or socks in a dormitory room. One then certainly can talk about the various possible ways these objects can be arranged. Of course, it then also is mathematically possible to define weighted sums over all these possible configurations, and one can thus formally define an "entropy," in mathematical formal analogy to thermodynamics. For example, one can define in this fashion a "dormitory entropy" for the arrangements of socks in a dormitory room. However, this "dormitory entropy" obviously has a very different meaning than physical entropy, and it would be foolish to conclude that useful work can be extracted from it, in the same fashion as would be possible for physical entropy. "Dormitory entropy" and physical entropy, while sharing the same mathematical formalism, have very different physical meaning. The same applies to information entropy, which simply does not have the same meaning as physical entropy.

6 Entropy and Energy

For physical entropy, there is a clear connection to physical energy. Physical entropy applies to systems with microscopic degrees of freedom subject to random thermal motion, such as gas molecules in a container. As gas particles bounce off the walls of their container, which might be static or which might be moving, these are physical processes that involve transfer of momentum and energy between the gas molecules and the walls. These microscopic bounces are the reason a gas volume exerts pressure on its walls, and this pressure entails work (physical energy) when the walls are moved. For example, if one moves the walls in, thereby reducing the gas' volume (phase space), real physical energy is required to do so, as work is required to move the wall against the gas pressure. This is the physical reason why a reduction of phase space, and thus entropy, requires expenditure of energy.

For information entropy, there is no such connection to physical energy. Bits, by their very nature, need to retain their value, unless—of course—they need to be changed according to the computation. In other words, bits do not perform random thermal "motion," and therefore they are very different from gas molecules. If bits were to change due to thermal fluctuations, they would not be bits any more. Moreover, it is rather obvious that a string of bits, such as a certain length of tape of a Turing machine, does not exert physical pressure on that tape. It is also rather obvious that changing the length of the tape of a Turing machine, and thus its information entropy, does not require physical work. In other words, changes in information entropy have no connection to changes in physical energy.

These points appear to be rather obvious, yet they are not appreciated by the *School of Dissipationless Computation.*

7 Thermodynamics Does Not Apply to Computation

As discussed in detail above, thermodynamics applies to a physical system where the microscopic degrees of freedom are subject to random thermal motion, and all that can be observed at the macroscopic level are averages (weighted sums). Thermodynamics does not apply to strings of bits since they cannot undergo random thermal motion, or they no longer would be bits. Thermodynamics does not apply to computation.

Computation requires the controlled deterministic switching of bits according to the rules of the algorithm to be executed. In a real physical system and in the presence of thermal noise, the bit operations and the bits themselves need to be isolated from thermal fluctuations in order to avoid random switching. This need to provide isolation from thermal noise is the fundamental reason for energy dissipation in computation, and not the logical irreversibility of bit operations.

Thermodynamics would apply to a string of bits, if these bits were to switch at random, but then they no longer would be useful for computation. In other words,

the more a system can be described by thermodynamics, the less it can be used for computation. (Just to make sure, this statement is true for the deterministic Turing-machine model used here. It does not apply to other models of computation, such as stochastic computing).

As has already been pointed out by us some time ago [5], there appears to be some "complementarity" between thermodynamics and computation. A system can either be described by thermodynamics, but then it cannot be used for computation. Or, if a system is to be used for computation, it cannot be described by thermodynamics.

We are led to the main conclusion of this chapter that the thermodynamics of computation is a contradiction in terms.

8 Experimental Studies of Landauer's Principle

In recent years, there have been several publications that claim to provide an experimental confirmation of Landauer's principle, i.e. that erasure is the true source of dissipation in a bit operation. What all of these studies have in common is to examine the switching process in a two-state system close to the noise floor. One such study [25] investigates a single colloidal particle in a modulated double-well potential realized by an optical trap. Another study [26] investigates, using magneto-optical measurements, the switching of nanoscale magnetic bits. A very recent study [27] claims to experimentally demonstrate a quantum version of Landauer's principle using a single atom in an ion trap.

While it is not in the scope of this chapter to refute each and every of these studies (and others) in detail, we note that they all set up a switching protocol, which they call "erasure." Studying the energetics of their such-defined switching process, and showing that this energy is close to k_BT, they then conclude that erasure requires dissipation. As such, these studies "conclude" their own assumptions. What these studies really do is to show that switching in a two-state system requires an energy of k_BT, which hardly is a surprising result. As we have argued before, there is nothing special about "erasure," which is a bit operation just like any other.

A true experimental demonstration of Landauer's contention that only erasure requires dissipation would be experiments that show that a bit operation called "erasure" is dissipative, and other non-erasure bit operations do not require dissipation. Of course, there is no such experiment (since it does not exist).

So, the problem with these experiments is not with the experiment itself, but with their interpretation. It is not surprising to find an energy limit on the order of k_BT when studying the energetics of switching in a two-state system. It certainly is not true that by experimentally demonstrating switching energies of k_BT one has provided a validation of Landauer's arguments.

Of course, when dealing with such small energies on the order of k_BT, there also is ample room for experimental errors. For example, [26] obtain the small switching energy as the difference of two large energies, obtained from their

experiments. Unless these two large energies are very well known, there is room for systematic errors when taking their difference. Their Fig. 4 shows their results for the switching energy as a function of temperature (in the range of 300–400 K), which—of course—should be linear in temperature, but the experimental data does not show a temperature dependence, raising the suspicion of systematic errors in these experiments. The same might be true for the other experiments, which deal with large energies, yet draw conclusions about very small energies.

There is also an experimental study that claims to provide an experimental test of Landauer's principle at the sub-k_BT level [28]. This study is different from the above experiments in that it does not deal with the switching in a two-state system. Rather, it studies the energetics of adiabatically charging and discharging a capacitor. We have already argued above that such adiabatic processes have nothing to do with thermodynamics and Landauer's arguments. While we disagree with the interpretation with this experiment, it is a nice demonstration of the possibility of sub-k_BT energy dissipation in adiabatically charging and discharging a capacitor.

9 Landauer and Maxwell's Demon

While not central to the main message of this chapter, Landauer and the *School of Dissipationless Computation* also have contributed to a re-interpretation of Maxwell's demon, which we shall briefly discuss here.

The connection between information and thermodynamics has a long history, going back to Boltzmann and Maxwell's demon [29]. In an early controversy regarding Boltzmann's then-young theory of thermodynamics, Maxwell conceived of a Gedankenexperiment to show that one could—in principle—violate the second law of thermodynamics. He argued that an intelligent "demon," who operates a (frictionless) door between two gas containers, could cool one container and heat the other by open that door just at the right time to let cool molecules pass one way, and hot molecules the other.

A resolution of this apparent paradox was given by Szilard, who argued that the demon had to "know" when to open the door and when to keep it closed. This required the demon to perform a measurement on an approaching gas molecule, and the acquisition of this information in a noisy environment required an amount of energy that restored the validity of thermodynamics. According to Szilard, it is the acquisition of information through a noisy channel that is the fundamental reason for the expenditure of energy.

Since the *School of Dissipationless Computation* believes that the fundamental reason for energy dissipation is the erasure of information and that measurements can be done without expenditure of energy, they re-interpreted Maxwell's demon in the following way: The demon can make the measurement without expending energy, but this information then is stored in the demon's mind. It is when the demon erases its mind that the energy dissipation occurs.

There are obvious problems with this interpretation. What if the demon chooses to clear its mind 1 year after the operations on the gas molecules? Of course, the energy needs to be expended when the gas molecules move, and not at some arbitrary later point in time.

However, such an obvious contradiction with this re-interpretation of Maxwell's demon does not matter since the arguments of the *School of Dissipationless Computation* are incorrect in the first place.

10 Summary and Conclusion

In this chapter, we have attempted a comprehensive discussion of Landauer's original arguments regarding logical irreversibility and heat generation in the computing process. In particular, we have attempted to carefully define what we mean by general terms, such as "computation," "reversibility," and "entropy." The literature contains quite a bit of confusion due to a lack of clear definitions of these terms. In particular, we have gone to great length to distinguish between physical (thermodynamic) entropy and information entropy. We found that Landauer's principle contains a fundamental flaw in using information entropy as if it were physical entropy. With that, the foundation for the *School of Dissipationless Computation* collapses.

We also discussed several experimental studies, which claim to provide experimental proof of Landauer's principle. We concluded that this is not so. While the experiments may be valid, their interpretation is not.

We discussed at great length that thermodynamics does not apply to computation. The main conclusion of this chapter is that the thermodynamics of computation is a contradiction in terms.

References

1. R.W. Keyes, R. Landauer, Minimal energy dissipation in logic. IBM J. Res. Dev. **14**, 152–157 (1970)
2. R.W. Keyes, Power dissipation in information processing. Science **168**, 796–801 (1970)
3. R. Landauer, Information is physical. Phys. Today **44**(5), 23–29 (1991)
4. R. Landauer, Irreversibility and heat generation in the computing process. IBM J. Res. Dev. **5**(3), 183–191 (1961)
5. W. Porod, R.O. Grondin, D.K. Ferry, G. Porod, Dissipation in computation. Phys. Rev. Lett. **52**(3), 232–235 (1984). Rebuttals by: C.H. Bennett, Phys. Rev. Lett. **52**(12), 1202 (1984); P. Benioff, ibid, p. 1203; T. Toffoli, ibid, p. 1204, and R. Landauer, ibid, p. 1205; Porod et al. Respond, ibid, p. 1206
6. W. Porod, Comment on 'energy requirement in communication'. Appl. Phys. Lett. **52**, 2191 (1988)
7. L.B. Kish, End of Moore's law: thermal (noise) death of integration in micro and nano electronics. Phys. Lett. A **305**, 144–149 (2002)

8. L.B. Kish, C.G. Granqvist, Energy requirement of control: comments on Szilard's engine and Maxwell's demon. Europhys. Lett. **98**, 68001 (2012)
9. L.B. Kish, C.G. Granqvist, S.P. Khatri, H. Wen, Demons: Maxwell's demon, Szilard's engine and Landauer's erasure–dissipation. Int. J. Mod. Phys. Conf. Ser. **33**, 1460364 (2014)
10. J.D. Norton, Eaters of the lotus: Landauer's principle and the return of Maxwell's demon. Stud. Hist. Philos. Sci. B **36**, 375–411 (2005)
11. J.D. Norton, Waiting for Landauer. Stud. Hist. Philos. Sci. B **42**, 184–198 (2011)
12. J.D. Norton, The end of the thermodynamics of computation: a no-go result. Philos. Sci. **80**(5), 1182–1192 (2013)
13. T. Sagawa, M. Ueda, Minimal energy cost for thermodynamic information processing: measurement and information erasure. Phys. Rev. Lett. **102**, 250602 (2009)
14. A. Turing, in 1948, Intelligent Machinery, ed. By C.R. Evans, A.D.J. Robertson. Reprinted in Cybernetics: Key Papers (University Park Press, Baltimore, 1968), p. 31
15. E. Fredkin, T. Toffoli, Conservative logic. Int. J. Theor. Phys. **21**(3/4), 219–253 (1982)
16. D.J. Frank, Comparison of High Speed Voltage-scaled Conventional and Adiabatic Circuits, in *Proc. Int. Workshop Low Power Electron. Design*, (IEEE, New York, 1996), pp. 377–380
17. P. Solomon, D.J. Frank, The Case for Reversible Computation, in *Proc. Int. Workshop Low Power Design*, (ACM, New York, 1994), pp. 93–98
18. M.P. Frank, Introduction to Reversible Computing: Motivation, Progress, and Challenges, in *Proceedings of the 2nd Conference on Computing Frontiers*, (ACM, New York, 2005), pp. 385–390
19. M.P. Frank, in *Foundations of Generalized Reversible Computing*, eds. By I. Phillips, H. Rahaman. *Reversible Computation. RC 2017. Lecture Notes in Computer Science*, vol. 10301 (Springer, Berlin, 2017), pp. 19–34
20. C.H. Bennett, Logical reversibility of computation. IBM J. Res. Dev. **17**(6), 525–532 (1973)
21. C.H. Bennett, The thermodynamics of computation – a review. Int. J. Theor. Phys. **21**(12), 905–940 (1982)
22. C.H. Bennett, Notes on the history of reversible computation. IBM J. Res. Dev. **44**(1/2), 270–277 (2000)
23. C.H. Bennett, Notes on Landauer's principle, reversible computation, and Maxwell's demon. Stud. Hist. Philos. Mod. Phys. **34**, 501–510 (2003)
24. E. Lutz, S. Ciliberto, Information: From Maxwell's demon to Landauer's eraser. Phys. Today **68**(9), 30 (2015)
25. A. Berut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, E. Lutz, Experimental verification of Landauer's principle linking information and thermodynamics. Nature **483**, 187–189 (2012)
26. J. Hong, B. Lambson, S. Dhuey, J. Bokor, Experimental test of Landauer's principle in single-bit operations on nanomagnetic memory bits. Sci. Adv. **2**, e1501492 (2016)
27. L.L. Yan, T.P. Xiong, K. Rehan, F. Zhou, D.F. Liang, L. Chen, J.Q. Zhang, W.L. Yang, Z.H. Ma, M. Feng, Single-atom demonstration of quantum Landauer principle, arXiv:1803.10424 [quant-ph] (2018)
28. A.O. Orlov, C.S. Lent, C.C. Thorpe, G.P. Boechler, G.L. Snider, Experimental test of Landauer's principle at the sub-k_BT level. Jpn. J. Appl. Phys. **51**, 06FE10 (2012)
29. H. Leff, A.F. Rex, *Maxwell Demon 2: Entropy, Classical and Quantum Information, Computing* (IOP Publishing, Bristol, 2002)

The Physics of Information: From Maxwell to Landauer

Sergio Ciliberto and Eric Lutz

Contents

1 Introduction

In 1991 Rolf Landauer argued that information is physical [1]. Since information is processed in physical devices, he concluded that information has to obey the laws of physics, and in particular the laws of thermodynamics. Information is thus stored in physical systems, such as books or memory sticks, and transmitted by physical means, for instance with the help of electrical or optical signals.

S. Ciliberto (✉)
Université de Lyon, CNRS, Laboratoire de Physique, École Normale Supérieure de Lyon (UMR5672), Lyon Cedex 07, France
e-mail: sergio.ciliberto@ens-lyon.fr

E. Lutz
Department of Physics, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany

C. S. Lent et al. (eds.), *Energy Limits in Computation*,
https://doi.org/10.1007/978-3-319-93458-7_5

But what is 'information'? A simple, intuitive answer is 'what you don't already know.' If someone tells you that the earth is spherical, you surely would not learn much: this message has low information content. However, if you are told that the oil price will double the day after tomorrow, assuming for a moment this to be true, you would learn a great deal : this message has hence high information content. Mathematically, the amount of information is quantified by the so-called information entropy H introduced by Claude Shannon in 1948; the larger the entropy, the bigger the information content [2]. The simplest device to store information is a system with two distinct states, for example up/down, left/right or magnetization/no magnetization. If the system is known to be with probability one in one of either states, probing the system will not reveal any new information, and the Shannon entropy is zero. On the other hand, if the two states can be occupied with probability one-half, and the actual state is therefore initially undetermined, an examination of the system will provide information about the state it is in. In this case, the Shannon entropy is equal to $\ln(2)$. This value corresponds to the smallest amount of information and is called a bit. A two-state system can thus store up to one bit of information.

The second law of thermodynamics, as formulated by Rudolf Clausius in 1850, is based on the empirical observation that some processes only occur spontaneously in one preferred direction [3]. Everyone who forgot a cup of hot tea on a table has noted that heat flows by itself from a hotter (the cup) to a colder body (the room), and never the other way around. Heat flow is therefore said to be irreversible. Clausius characterized the irreversibility of natural macroscopic processes by defining the thermodynamic entropy S, a quantity that is not conserved, in contrast to energy, but can only increase in isolated systems. This asymmetry in the change of entropy imposes restrictions on the type of physical phenomena that are possible. Similarly, the application of the second law of thermodynamics to information sets limitations on information processing tasks such as transmission or erasure. More general questions address the thermodynamic consequences of information gain. In particular, whether it is possible to extract useful mechanical work from a system by observing its state, and if yes how much. And at the more fundamental level: are thermodynamic and information entropies related [4, 5]?

1.1 Maxwell's Demon and Szilard's Engine

The first hint of a connection between information and thermodynamics may be traced back to James Clerk Maxwell's now famous demon introduced in 1867 [6–8]. The demon is an intelligent creature able to monitor individual molecules of a gas contained in two neighboring chambers initially at the same temperature, as shown in Fig. 1. The temperature of the gas is defined by the mean kinetic energy of the molecules and is hence proportional to their mean-square velocity. However, not all the particles will have the same velocity. Some of the molecules will be going faster than average and some will be going slower. By opening and closing a molecular-sized trap door in the partitioning wall, the demon collects the faster molecules in

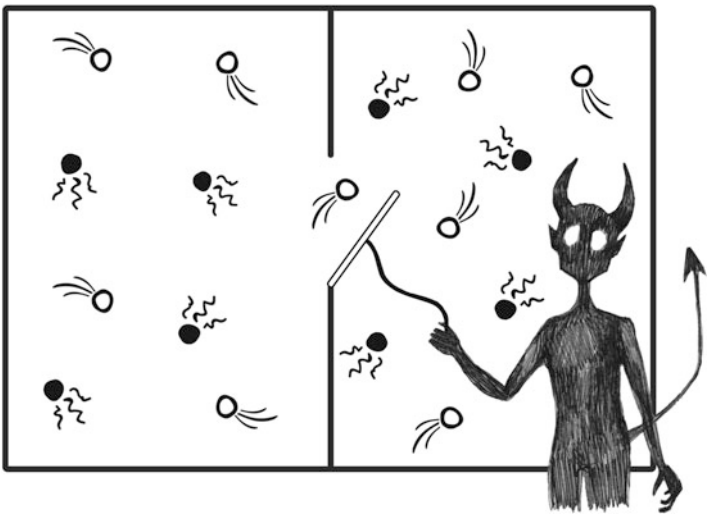

Fig. 1 Maxwell's demon. By detecting the positions and velocities of gas molecules in two neighboring chambers and using that information to time the opening and closing of a trapdoor that separates them, a tiny, intelligent being could, in theory, sort molecules by velocity. By doing so, it could create a temperature difference across the chambers that could be used to perform mechanical work. If the trapdoor is frictionless, the sorting requires no work from the demon himself, in apparent violation of the second law of thermodynamics (drawn by Claire Lebeau)

one of the chambers and the slower ones in the other. The two chambers now contain gases with different mean-square velocities and hence different temperatures. This temperature difference may be used to run a heat engine and produce mechanical work. By gathering information about the position and velocity of each particle and using this knowledge to sort them, the demon is therefore able to decrease the entropy of the system and convert the acquired information into energy. The problem is that the demon, assuming a frictionless trap door, is able to do all this without performing any work himself, in apparent violation of the second law of thermodynamics. The proper resolution of this paradox took 115 years.

A simplified one-particle engine has been suggested by Leo Szilard in 1929 [9]. In this setup, schematically shown in Fig. 2, the gas consists of a single molecule and the wall separating the identical chambers is replaced by a moving piston to which a weight can be attached. We now have a two-state system very similar to the one discussed above. Initially, the particle has a probability of one half to be in one of the two chambers. By looking into the container the demon acquires information about the actual state of the system, learning what he did not know before. If the molecule is found in the right chamber, the weight is attached to the right-hand side of the piston which is then released from its former position. During the expansion of the gas, the piston is pushed to the left and the weight is pulled upwards, performing work against gravity. The piston is attached to the left-hand side of the piston when the molecule is observed in the left chamber. The second law of thermodynamics

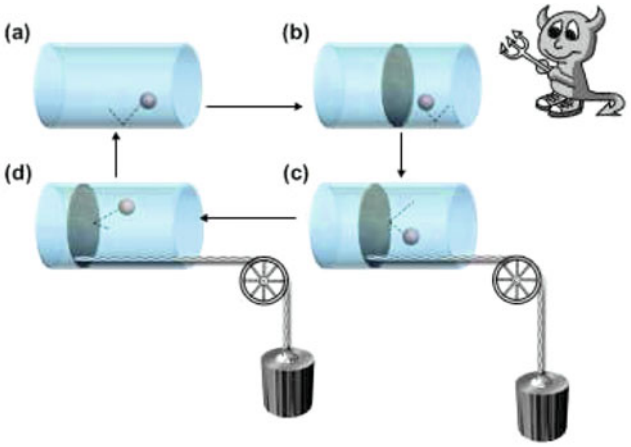

Fig. 2 Szilard's engine. A crafty observer can turn a single particle in a box into an engine that converts information into mechanical work. If, say, (**a**) the particle is found on the box's left-hand side, (**b**) the observer inserts a movable wall and (**c**) attaches a weight to its left side; (**d**) the free expansion of the one-particle gas pushes the wall to the right, lifts the weight, and thereby performs work against gravity (adapted from Ref. [8])

limits the maximum amount of work that can be produced by the Szilard engine to $k_B T \ln(2)$, where k_B is the Boltzmann constant and T the temperature of the gas. This corresponds to the maximum amount of energy that can be obtained by converting one bit of information, and is historically the first clear statement of the relationship between information and energy. In modern language, this result further implies that information and thermodynamic entropies are equal, $S = k_B H$, up to the multiplicative factor k_B introduced for dimensional reasons (the Shannon entropy H is dimensionless).

1.2 Landauer's Principle and Bennett's Resolution

It is useful to distinguish two complementary aspects: the first one is information gain, as we have just discussed with Maxwell's demon, the second one is information erasure, which has been investigated from a thermodynamic point of view by Landauer in 1961. Let us again consider a two-state system and let us assume that it initially stores one bit of information, that is, the two states are occupied with equal probability one-half. This bit may be erased by resetting the system to one of the states, which will then be occupied with unit probability, a situation that corresponds to a zero Shannon entropy. By applying the second law of thermodynamics, Landauer demonstrated that information erasure is necessarily a dissipative process: the erasure of one bit of information is accompanied by the production of at least $k_B T \ln(2)$ of heat into the environment. This result is known as Landauer's erasure principle. It emphasizes the fundamental difference

between the process of writing and erasing information. Writing is akin to copying information from one device to another: state left is mapped to left and state right is mapped to right, for example. This one-to-one mapping can be realized in principle without dissipating any heat (in statistical mechanics one would say that it conserves the volume in phase space). By contrast, erasing information is a two-to-one transformation: states left and right are mapped onto one single state, say right (this process does not conserve the volume in phase space and is thus dissipative).

Landauer's principle played a central role in solving the paradox of Maxwell's demon. In 1982 Charles Bennett noted that the demon has to store the information he acquires about the gas molecules in a memory [10]. After a full information gathering energy producing cycle, this memory has to be reset to its initial state to allow for a new iteration, and its information content has thus to be erased (a similar argument was put forward by Oliver Penrose in 1970 [11]). According to Landauer's principle, the erasure process will dissipate an amount of energy that is always larger than the quantity of energy produced by the demon during one cycle. The demon has consequently to pay an energetic price to sort the molecules and have heat flow from the colder chamber to the hotter chamber, in full agreement with the second law of thermodynamics. Before Bennett's resolution, it was often believed, following arguments put forward by Leon Brillouin and Dennis Gabor, that it was the energetic price of the measurement, that is, of the act of gathering information, that would save the second law [12]. However, as shown by Bennett, there is no fundamental energetic limitation on the measurement process, which like the copy operation may in principle be performed without dissipation, in stark contrast to erasure.

Box 1 Landauer's Erasure Principle

Landauer's principle can be seen as a direct consequence of the second law of thermodynamics. Consider a system (SYS) coupled to a reservoir (RES) at temperature T. According to the second law, the total entropy change for system and reservoir is positive: $S_{\mathrm{TOT}} = S_{\mathrm{SYS}} + S_{\mathrm{RES}} \geq 0$. Since the reservoir is always at equilibrium, owing to its very large size, we have followed Clausius, $\Delta S_{\mathrm{RES}} = Q_{\mathrm{RES}}/T$. In other words, the heat absorbed by the reservoir satisfies $Q_{\mathrm{RES}} \geq T\Delta S_{\mathrm{SYS}}$. For a two-state system that stores one bit of information, there are initially two possible states that can be occupied with probability one half, and the initial Shannon entropy is $H_i = \ln(2)$. After erasure, the system is with unit probability in one of the states and the final Shannon entropy vanishes $H_f = 0$. The change of information entropy is thus $\Delta H = -\ln(2)$. During this erasure process the ability of the system to store information has been modified. By further using the (assumed) equivalence between thermodynamic entropy S and information entropy H we can write $\Delta S_{\mathrm{SYS}} = k_{\mathrm{B}} H = k_{\mathrm{B}} \ln(2)$. We hence obtain $Q_{\mathrm{RES}} \geq k_{\mathrm{B}} T \ln(2)$, showing that the heat dissipated into the reservoir during the erasure of one bit of information is always larger than $k_{\mathrm{B}} T \ln(2)$.

2 Experimental Implementations

For almost a century and a half, the demon belonged to the realm of a gedanken experiment as the tracking and manipulation of individual microscopic particles was impossible. However, owing to the remarkable progress achieved in the last decades, such experiments have now become feasible. Just to give a hint on what can be done, we will discuss in the following sections several experimental realizations of Maxwell's demon and Szilard's engine, as well as several verification of Landauer's principle.

2.1 Experiments on Maxwell's Demon

The first realization of a Maxwell demon was used to cool atoms in a magnetic trap. An ensemble of atoms is first trapped in a magnetic trap (see Fig. 3) [14]. A one way barrier (which plays the role of the demon) sweeps the magnetic trap from the right to the left, starting at a very large value of the potential. The atoms reaching this position have transformed almost all their kinetic energy in potential energy and are, therefore, very cool. These atoms go through the barrier but they cannot come back, i.e. the barrier behaves as an atom-diode [13–15]. Thus the hot atoms are on the right and the cold atoms are on the left. At the end of the process when the sweeping one-way barrier reaches the bottom of the magnetic potential all of the atoms are cooled down. The one way barrier is composed by two laser beams suitable tuned to atomic transitions. In Fig. 3 (left) one of the two lasers is on the left of the barrier and forces the atoms in an excited state. The frequency of the second laser, which is on the right of the barrier, is tuned in such a way that it has no effect on the atoms in the excited state and it repels the atoms in the ground state. Thus the atoms coming from the right, which are prepared in the excited state, go through the barrier and relax to the ground state by emitting a photon. Instead the atoms coming from the left, which are in the ground state, encounter first the barrier and remain trapped because they are repelled. Where does the connection with Maxwell demon come from? Indeed each time that an atom loses a photon the entropy of the light shining the atoms increases because before all the photons were coherently in the laser beam (low entropy state) and now the emitted photons are scattered in all directions (high entropy state). This entropy is related to an information entropy because each time that a photon is emitted we know that an atom has been cooled. It can be shown that indeed this gain of entropy is larger than the reduction of entropy produced by the cooling of the atomic cloud. It is important to notice that in this example the demon has not to be an intelligent being but it is just a suitable tuned device which automatically implements the operation.

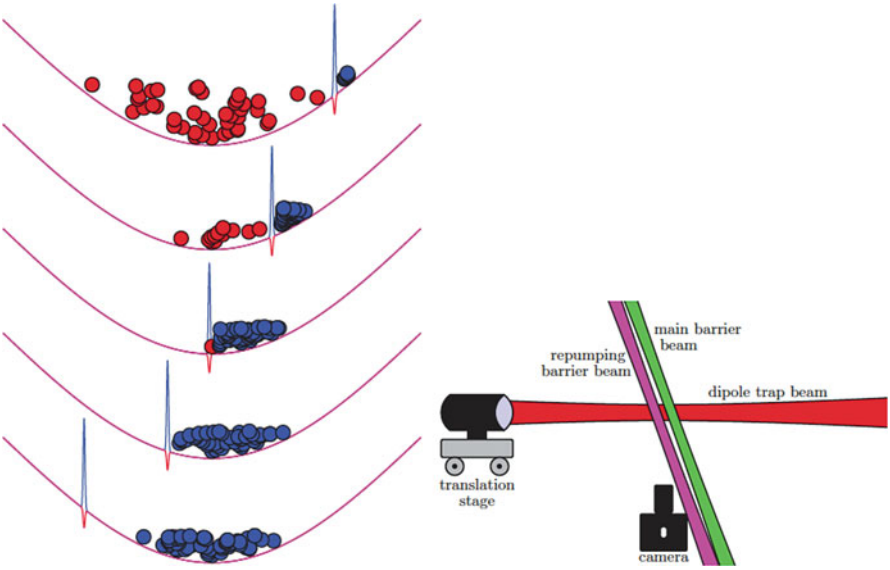

Fig. 3 Using a Maxwell's demon to cool atoms. A pair of laser beams can be tuned to atomic transitions and configured to create a one-way potential barrier; atoms may cross unimpeded in one direction, from left to right left in this figure, but not in the other. Left panel : when the barrier is introduced at the periphery of the trapping potential, (right side) the atoms that cross the barrier will be those that have converted nearly all their kinetic energy to potential energy, in other words, the cold ones. By slowly sweeping the barrier (from the right to the left) across the trapping potential, one can sort cold atoms (blue) from hot ones (red), reminiscent of Maxwell's famous thought experiment, or cool an entire atomic ensemble. Because the cold atoms do work against the optical barrier as it moves, their kinetic energy remains small even as they return to the deep portion of the potential well. Right panel: schematic representation of the optical set-up showing the optical trap (red beam), the translational stage and the two beams one way barrier (adapted from Ref. [13])

2.1.1 The Szilard Engine: Work Production from Information

A Szilard engine has been realized in 2010 by using a single microscopic Brownian particle in a fluid and confined to a spiral-staircase-like potential shown in Fig. 4 [16]. Driven by thermal fluctuations, the particle performs an erratic up and down motion along the staircase. However, because of the potential gradient downwards steps will be more frequent than upwards steps and the particle will on average fall down. The position of the particle is measured with the help of a CCD camera. Each time the particle is observed to jump upwards, this information is used to insert a potential barrier that hinders the particle to move down. By repeating this procedure, the average particle motion is now upstairs and work is done against the potential gradient. By lifting the particle mechanical work has therefore been produced by gathering information about its position. This is the first example of a device that converts information into energy for a system coupled to a single thermal

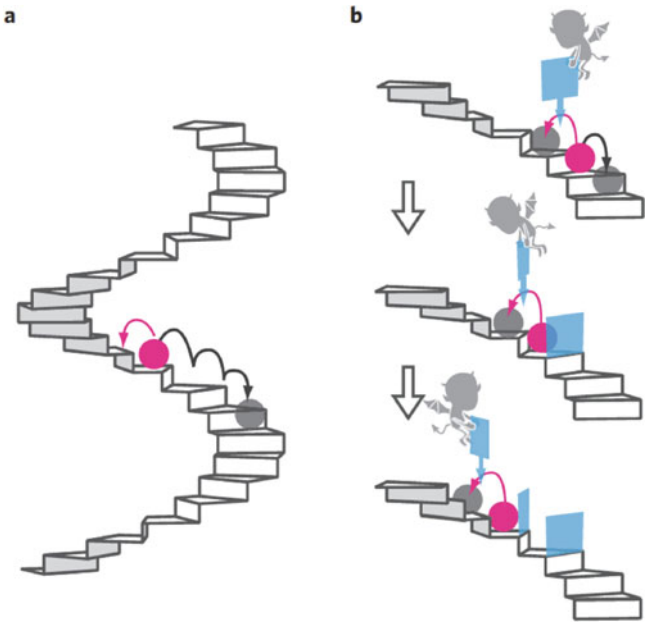

Fig. 4 Experimental realization of Szilard's engine. (**a**) A colloidal particle in a staircase potential moves downwards on average, but energy fluctuations can push it upwards from time to time. (**b**) When the demon observes such an event, he inserts a wall to prevent downward steps. By repeating this procedure, the particle can be brought to move upwards, performing work against the force created by the staircase potential. In the actual experiment, the staircase potential is implemented by a tilted periodic potential and the insertion of the wall is simply realized by switching the potential, replacing a minimum (no wall) by a maximum (wall) (adapted from Ref. [16])

environment. However there is not a contradiction with the second law because Sagawa and Ueda formalized the idea that information gained through microlevel measurements can be used to extract added work from a heat engine [17]. Their formula for the maximum extractable work is:

$$\langle W_{\max} \rangle = -\Delta F + k_{\mathrm{B}}\, T\, \langle I \rangle \tag{1}$$

where ΔF is the free energy difference between the final and initial state and the extra term represents the so-called mutual information I. In absence of measurement errors this quantity reduces to the Shannon entropy : $I = -\sum_k P(\Gamma_k) \ln[P(\Gamma_k)]$, where $P(\Gamma_m)$ is the probability of finding the system in the state Γ_k. Then in the specific case of the previously described staircase potential [16]: $I = -p \ln p - (1-p) \ln p$ where p is the probability of finding the particle in a specific region.

In this context the Jarzynski equality (see "Appendix 1: Stochastic Thermodynamics and Information Energy Cost") also contains this extra term and it becomes:

$$\langle \exp(-\beta W + I) \rangle = \exp(-\beta \Delta F) \tag{2}$$

which leads to

$$\langle W \rangle \geq \Delta F - k_B T \langle I \rangle \tag{3}$$

Equations (2) and (3) generalize the second law of thermodynamics taking into account the amount of information introduced into the system [5, 18]. Indeed Eq. (3) indicates that thanks to information the work performed on the system to drive it between an initial and a final equilibrium states can be smaller than the free energy difference between the two states. Equation (2) has been directly tested in a single electron transistor [19].

2.1.2 The Autonomous Maxwell Demon Improves Cooling

An autonomous Maxwell demon using a local feedback mechanism allows an efficient cooling of the system [20, 21]. The device, whose principle is sketched in Fig. 5a, is composed by a SET (Single Electron Transistor) formed by a small normal metallic island connected to two normal metallic leads by tunnel junctions, which permit electron transport between the leads and the island. The SET is biased by a potential V and a gate voltage V_g, applied to the island via a capacitance, controls the current I_e flowing through the SET. The island is coupled capacitively with a single electron box which acts as a demon which detects the presence of an electron in the island and applies a feedback. Specifically when an electron tunnels to the island, the demon traps it with a positive charge (panels 1 and 2). Conversely, when an electron leaves the island, the demon applies a negative charge to repel further electrons that would enter the island (panels 3 and 4). This effect is obtained by designing the electrodes of the demon in such a way that when an electron enters the island from a source electrode, an electron tunnels out of the demon island as a response, exploiting the mutual Coulomb repulsion between the two electrons. Similarly, when an electron enters to the drain electrode from the system island, an electron tunnels back to the demon island, attracted by the overall positive charge. The cycle of these interactions between the two devices realizes the autonomous demon, which allows the cooling of the leads. In the experimental realization presented in [20], the leads and the demon were thermally insulated, and the measurements of their temperatures is used to characterize the effect of the demon on the device operation. In Fig. 5b we plot the variation of the leads temperatures as a function of $n_g \propto V_g$ when the demon acts on the system. We clearly see that around $n_g = 1/2$ the two leads are both cooled of 1 mK at a mean temperature of 50 mK. This occurs because the tunneling electrons have to take the energy from the thermal energy of the leads, which, being thermally isolated, cool

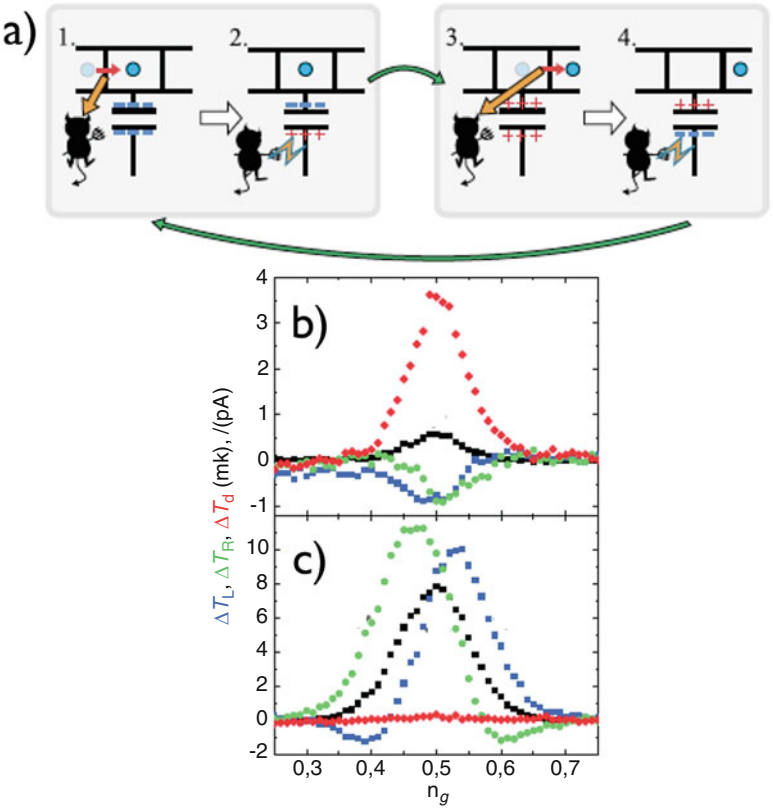

Fig. 5 (**a**) Principle of the experimental realization of the autonomous Maxwell demon. The horizontal top row schematizes a Single Electron Transistor. Electrons (blue circle) can tunnel inside the central island from the left wall and outside from the right wall. The demon watches at the state of the island and it applies a positive charge to attract the electrons when they tunnel inside and they repels them when they tunnel outside. The systems cools because of the energy released toward the heat bath by the tunneling events and the presence of the demon makes the cooling processes more efficient. The energy variation of the processes is negative because of the information introduced by the demon. (**b**) The measured temperature variations of the left (blue line) and right (green line) leads as a function of the external control parameter n_g when the demon is active and the bath temperature is 50 mK. We see that at the optimum value $n_g = 1/2$ both leads are cooled of about 1 mK and the current I_e flowing through the SET (black line) has a maximum. At the same time in order to processes information the temperature of the demon (red line) increases of a few mK. (**c**) The same parameter of the panel (**b**) are measured when the demon is not active. We see that the demon temperature does not change, whereas both leads are now heated by the current I_e (adapted from Ref. [20])

down. This increases the rate at which electrons tunnel against Coulomb repulsion, giving rise to increased cooling power. At the same time the demon increases its temperature because it has to dissipate energy in order to process information, as discussed in Ref. [22]. Thus the total (system+demon) energy production is positive.

The coupling of the demon with the SET can be controlled by a second gate which acts on the single electron box. In Fig. 5c we plot the measured temperatures when the demon has been switched off. We clearly see that in such a case the demon temperature does not change and the two electrodes are heating up because of the current flow. This is the only example which shows that under specific conditions an autonomous local Maxwell demon, which does not use the external feedback, can be realized.

2.2 Experiments on Landauer's Principle

The experiments in the last section show that one can extract work from information. In the rest of this section we will discuss the reverse process, i.e. the energy needed to erase information. By applying the second law of thermodynamics, Landauer demonstrated that information erasure is necessarily a dissipative process: the erasure of one bit of information is accompanied by the production of at least $k_B T \ln(2)$ of heat into the environment. This result is known as Landauer's erasure principle. It emphasizes the fundamental difference between the process of writing and erasing information. Writing is akin to copying information from one device to another: state left is mapped to left and state right is mapped to right, for example. This one-to-one mapping can be realized in principle without dissipating any heat (in statistical mechanics one would say that it conserves the volume in phase space). By contrast, erasing information is a two-to-one transformation: states left and right are mapped onto one single state, say right (this process does not conserve the volume in phase space and is thus dissipative).

Landauer's original thought experiment has been realized for the first time in a real system in 2011 using a colloidal Brownian particle in a fluid trapped in a double-well potential produced by two strongly focused laser beams [23, 24]. This system has two distinct states (particle in the right or left well) and may thus be used to store one bit of information. The erasure principle has been verified by implementing a protocol proposed by Bennett and illustrated in Fig. 6. At the beginning of the erasure process, the colloidal particle may be either in the left or right well with equal probability of one half. The erasure protocol is composed of the following steps: (1) the barrier height is first decreased by varying the laser intensity, (2) the particle is then pushed to the right by gently inclining the potential and (3) the potential is brought back to its initial shape. At the end of the process, the particle is in the right well with unit probability, irrespective of its departure position. As in the previous experiment, the position of the particle is recorded with the help of a camera. For a full erasure cycle, the average heat dissipated into the environment is equal to the average work needed to modulate the form of the double-well potential. This quantity was evaluated from the measured trajectory and shown to be always larger than the Landauer bound which is asymptotically approaches in the limit of long erasure times. However, in order to reach the bound, the protocol must be accurately chosen because as discussed in Ref. [23] and shown

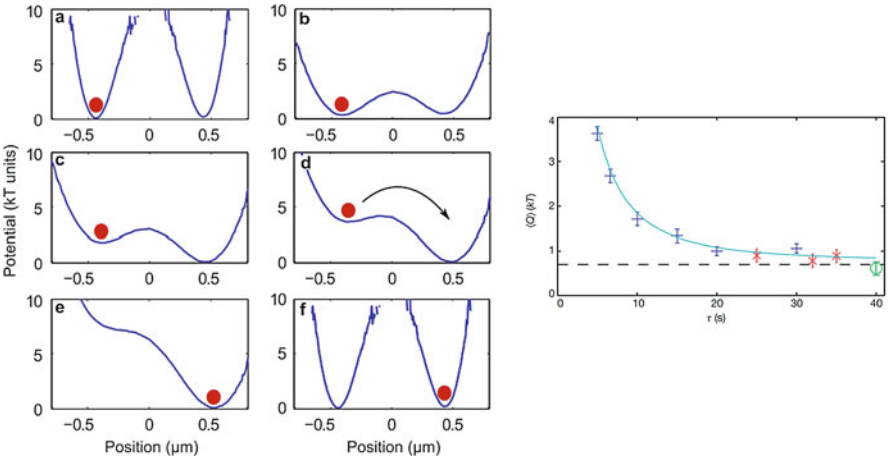

Fig. 6 Experimental verification of Landauer's erasure principle. (**a**) A colloidal particle is initially confined in one of two wells of a double-well potential with probability one-half. This configuration stores one bit of information. (**b**) By modulating the height of the barrier and (**c**) applying a tilt, (**d**) the particle can be brought to one of the wells with probability one, irrespective of the initial position. This final configuration corresponds to zero bit of information. In the limit of long erasure cycles, the heat dissipated during the erasure process can approach, but not exceed, the Landauer bound indicated by the dashed line (adapted from Ref. [23])

experimentally [25] there are protocols that are intrinsically irreversible no matter how slow are performed. The way in which a protocol can be optimized has been theoretically solved in Ref. [26] but the optimal protocol is not often easy to apply in an experiment.

2.3 Other Experiments on the Physics of Information

By having successfully turned gedanken into real experiments, the above four seminal examples provide a firm empirical foundation to the physics of information and the intimate connection existing between information and energy. This connection is reenforced by the relationship between the generalized Jarzinsky equality [27] and the Landauer bound which has been proved and tested on experimental data in Ref. [24] and shortly summarized in the "Appendix 1: Stochastic Thermodynamics and Information Energy Cost" of this chapter .

A number of additional experiments have verified the erasure principle in various systems [28–34]. The latter include an electrical RC circuit [28] and a feedback trap [30, 31]. In addition, Ref. [32] has studied the symmetry breaking, induced in the probability distribution of the position of a Brownian particle, by commuting the trapping potential from a single to a double well potential. The authors measured the time evolution of the system entropy and showed how to produce work from

information. Finally, experiments on the Landauer bound have been performed in nano devices, most notably using a single electron box [29] and nanomagnets [33, 34]. These experiments open the way to insightful applications for future developments of information technology.

3 Extensions to the Quantum Regime

3.1 Experiments on Quantum Maxwell's Demon

The experimental investigation of the physics of information has lately been extended to the quantum regime. The group of Roberto Serra in Sao Paulo has successfully realized a quantum Maxwell demon in a Nuclear Magnetic Resonance (NMR) setup [35]. The demon was implemented as a spin-1/2 quantum memory that acquires information about another spin-1/2 system and employs it to control its dynamics. Using a coherent measured-based feedback protocol, the demon was shown to rectify the nonequilibrium entropy production due to quantum fluctuations and produce useful work. Concretely, the demon gained information about the system via a complete projective measurement. Based on the outcome of this measurement, a controlled evolution was applied to the system to balance the entropy production. Using quantum state tomography to reconstruct the density matrix ρ of the system at all times, the produced average work $\langle W \rangle$, or equivalently the mean entropy production $\langle \Sigma \rangle = \beta(\langle W \rangle - \Delta F)$, was shown to be bounded by the information gain, $\langle \Sigma \rangle \leq I_{\text{gain}}$. The latter quantifies the average information that the demon obtains by reading the outcomes of the measurement and is defined as $I_{\text{gain}} = S(\rho) - \sum_i p_i S(\rho_i)$, where ρ_i is the state after a measurement which occurs with probability p_i (see Fig. 7).

More recently, a quantum Maxwell demon has been implemented in a circuit QED system [36]. Here, the demon was a microwave cavity that encodes quantum

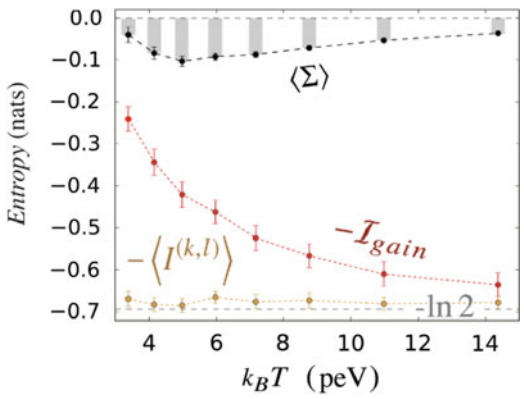

Fig. 7 Thermodynamics of a quantum Maxwell demon. Verification of the second law for the nonequilibrium mean entropy production, $\langle \Sigma \rangle = \beta(\langle W \rangle - \Delta F) \leq I_{\text{gain}}$, in the presence of quantum feedback as a function of temperature. The parameter I_{gain} quantifies the information gained through the measurement (adapted from Ref. [35])

information about a superconducting qubit and converts that information into work by powering up a propagating microwave pulse by stimulated emission. The power extracted from the system was directly accessed by measuring the difference between incoming and outcoming photons of the cavity. Using full tomography of the system, the entropy remaining in the demon's memory was further quantified and was shown to be always higher that the system entropy decrease, in agreement with the second law.

In addition in a quantum demon setting a multi-photon optical interferometer allowed the measure of the extractable work which was used as a thermodynamic separability criterion to assess the entanglement of two-qubit and three-qubit systems [37]. An experimental analysis of two-qubit Bell states and three-qubit GHZ and W states has confirmed that more work can be extracted from an entangled state than from a separable state. Bounds on the extractable work can therefore be employed as a useful thermodynamic entanglement witness.

3.2 Experiments on Quantum Landauer's Principle

Erasure of information encoded in quantum states has been first theoretically considered by Lubkin [39] and Vedral [40] (see also Ref. [8]). An experimental verification of the Landauer principle in a quantum setting has been recently reported using a molecular nanomagnet at a temperature of 1 K [38]. One bit of information was initially stored in a double-well potential of collective giant spin $S_z = \pm 10$ of a Fe_8 molecule. Work for the application of the tilt induced by a transverse magnetic field was determined via measurements of the magnetic susceptibility. Contrary to classical erasure which is achieved by decreasing the barrier height, here erasure was promoted by a thermally activated quantum tunnelling process. As a result, full erasure can be achieved much faster than in the classical regime. Using the product of the erasure work and the relaxation time, $W \cdot \tau_{\rm rel}$, as a figure of merit for the energy-time cost of information erasure, this experiment has reached the lower value to date with $W \cdot \tau_{\rm rel} \simeq 2 \times 10^{-23}$ erg/bit, as compared to 10^{-12} erg/bit. s for the classical experiment with the colloidal particle [23]. This puts the experiment close to the fundamental limit imposed by the Heisenberg uncertainty relation (see Fig. 8).

4 Applications

Landauer's principle applies not only to information erasure but also to all logically irreversible devices that possess more outputs than inputs. Thus, any Boolean gate operation that maps several input states onto the same output state, such as AND, NAND, and OR, has several states which are logically irreversible and will lead to the dissipation of an amount of heat of $k_{\rm B}T\ln(2)$ per processed bit, akin to

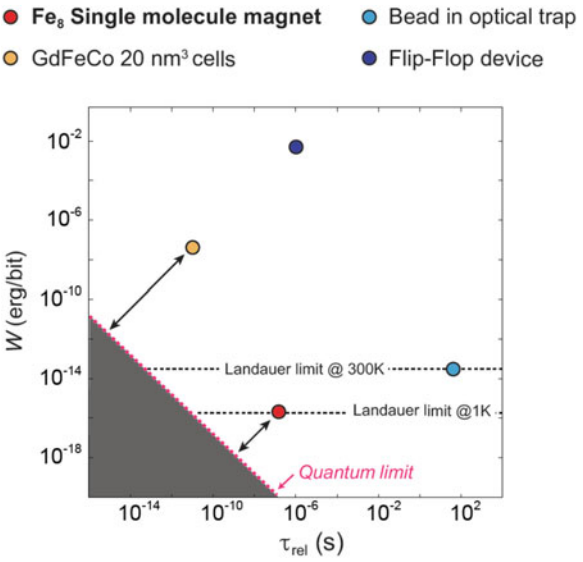

Fig. 8 Energy-time cost of erasure. The diagram shows the product of the energy and the time needed for erasure, $W \cdot \tau_{\mathrm{rel}}$, for various systems. The quantum limit is given by the Heisenberg uncertainty relation, $E \cdot \Delta t \geq \pi\hbar/2$. The Fe_8 molecule is currently the closest to the quantum limit (red dot) (adapted from Ref. [38])

the erasure process. As a result, Landauer's principle has important technological consequences. Heating laptops are nowadays becoming part of everyday experience. Heat production in microprocessors used in modern computers is known to be a major factor hindering their miniaturization, as it gets more and more difficult to evacuate excess heat when size, and thus surface, is reduced. While the overall heat dissipated in microchips is steadily decreasing, it still several orders of magnitude larger than the Landauer limit. However, the switching energy of a CMOS/FET transistor is predicted to reach the Landauer bound by 2035, indicating that engineers will soon face a fundamental physical limitation imposed by the second law of thermodynamics [41, 42]. This is remarkable as $k_{\mathrm{B}}T\ln(2)$ is about 3.10^{-21} Joule at room temperature and hence 22 orders of magnitude smaller than typical energy dissipated on our macroscopic scale. Recently, an experiment has demonstrated that Maxwell's demon can generate electric current and power by rectifying individual randomly moving electrons in small transistors [43].

Man-made computers are not the only existing information processing devices. Scientists have long realized that living biological cells can be viewed as biochemical information processors that may even outperform our current technology [44]. Cells are, for example, able to reproduce and create copies of themselves, acquire and process information coming from external stimuli, as well as communicate and exchange information with other cells. Recently, Landauer's principle has been employed to evaluate the energetic cost of a living cell computing the steady-state concentration of a chemical ligand in its surrounding environment [45]; it has been argued that it sets strong constraints on the design of cellular computing networks, as there is a tradeoff between the information processing capability of such a network and its energetic cost. Another important problem is the investigation of

ultrasensitive switches in molecular biology. A concrete example is the flagellar motor of E. coli bacteria that switches from clockwise to counterclockwise rotation depending on the intracellular concentration of a regulator protein. Switching mechanisms are highly complex and not fully understood. A mathematical framework that models the sensing of the protein concentration by the flagellar motor as a Maxwell demon has been successfully developed to calculate the rate of energy consumption needed to both sense and switch, and provide a quantitative description of the switching statistics [46]. More recent work has focused on the efficiency of cellular information processing [47], biochemical signal transduction [48], as well as on cost and precision of Brownian clocks [49] and computational copying in biochemical systems [50].

Maxwell's demon is therefore still vibrant 150 years after its inception. Together with Landauer's principle, he continues to play a prominent role in modern research as illustrated by the last examples. Having only very recently become an experimental science, information physics appears to have a promising future ahead.

Appendix 1: Stochastic Thermodynamics and Information Energy Cost

When the size of a system is reduced the role of fluctuations (either quantum or thermal) increases. Thus thermodynamic quantities such as internal energy, work, heat, and entropy cannot be characterized only by their mean values but also their fluctuations and probability distributions become relevant and useful to make predictions on a small system. Let us consider a simple example such as the motion of a Brownian particle subjected to a constant external force. Because of thermal fluctuations, the work performed on the particle by this force per unit time, i.e., the injected power, fluctuates and the smaller the force, the larger is the importance of power fluctuations [51–53]. The goal of stochastic thermodynamics is just that of studying the statistical properties of the above-mentioned fluctuating thermodynamic quantities in systems driven out of equilibrium by external forces, temperature differences, and chemical reactions. For this reason it has received in the last twenty years an increasing interest for its applications in microscopic devices, biological systems and for its connections with information theory[51–53].

Specifically it can be shown that the fluctuations on a time scale τ of the internal energy ΔU_τ, the work W_τ and the heat Q_τ are related by a first principle like equation, i.e.

$$\Delta U_\tau = U(t+\tau) - U(t) = \tilde{W}_\tau - Q_\tau \tag{4}$$

at any time t.

Furthermore the statistical properties of energy and entropy fluctuations are constrained by fluctuations theorems which impose bounds on their probability distributions (for more details see Ref. [51–53]). We summarize in the next section one of them which can be related to information and to Landuer's bound.

Estimate the Free Energy Difference from Work Fluctuations

In 1997 [54, 55] Jarzynski derived an equality which relates the free energy difference of a system in contact with a heat reservoir to the pdf of the work performed on the system to drive it from A to B along any path γ in the system parameter space. Specifically, when a system parameter λ is varied from time $t = 0$ to $t = t_s$, Jarzynski defines for one realization of the "switching process" from A to B the work performed on the system as

$$W = \int_0^{t_s} \dot{\lambda} \, \frac{\partial H_\lambda[z(t)]}{\partial \lambda} \mathrm{d}t, \tag{5}$$

where z denotes the phase-space point of the system and H_λ its λ-parametrized Hamiltonian.[1] One can consider an ensemble of realizations of this "switching process" with initial conditions all starting in the same initial equilibrium state. Then W may be computed for each trajectory in the ensemble. The Jarzynski equality states that [54, 55]

$$\exp\left(-\beta \Delta F\right) = \langle \exp\left(-\beta W\right) \rangle, \tag{6}$$

where $\langle \cdot \rangle$ denotes the ensemble average, $\beta^{-1} = k_\mathrm{B} T$ with k_B the Boltzmann constant and T the temperature. In other words $\langle \exp\left[-\beta W_\mathrm{diss}\right] \rangle = 1$, since we can always write $W = \Delta F + W_\mathrm{diss}$ where W_diss is the dissipated work. Thus it is easy to see that there must exist some paths γ such that $W_\mathrm{diss} \leq 0$. Moreover, the inequality $\langle \exp x \rangle \geq \exp \langle x \rangle$ allows us to recover the second principle, namely $\langle W_\mathrm{diss} \rangle \geq 0$, i.e. $\langle W \rangle \geq \Delta F$.

Landauer Bound and the Jarzynski Equality

We discuss in this appendix the strong relationship between the Jarzynski equality and the Landauer's bound. In Box 1 we presented the Landauer's principle as related to the system entropy. Let us consider as a specific example the experiment on

[1] This is a more general definition of work and it coincides with the standard one only if λ is a displacement (for more details see Ref. [53]).

the colloidal particle described in Sect. 2.2 [24]. In the memory erasure procedure which forces the system in the state 0, the entropy difference between the final and initial state is $\Delta S = -k_B \ln(2)$. In contrast the internal energy is unchanged by the protocol. Thus it is natural to await $\Delta F = k_B T \ln(2)$. However the ΔF that appears in the Jarzynski equality is the difference between the free energy of the system in the initial state (which is at equilibrium) and the equilibrium state corresponding to the final value of the control parameter: $F(\lambda(\tau)) - F(\lambda(0))$. Since the height of the barrier is always finite there is no change in the equilibrium free energy of the system between the beginning and the end of the procedure. Then $\Delta F = 0$, which implies $\left\langle e^{-\beta W_{st}} \right\rangle = 1$. Thus it seems that there is a problem between the Landauer principle (see Box 1) and the Jarzynski equality of Eq. (6).

Nevertheless Vaikuntanathan and Jarzyski [27] have shown that when there is a difference between the actual state of the system (described by the phase-space density ρ_t) and the equilibrium state (described by ρ_t^{eq}), the Jarzynski equality can be modified:

$$\left\langle e^{-\beta W_{st}(t)} \right\rangle_{(x,t)} = \frac{\rho^{eq}(x, \lambda(t))}{\rho(x, t)} e^{-\beta \Delta F(t)} \tag{7}$$

where $\langle . \rangle_{(x,t)}$ is the mean on all the trajectories that pass through x at time t.

In the experiment presented in Sect. 2.2, the selection of the trajectories where the information is actually erased corresponds to fix x to the chosen final well at the time $t = \tau$. It follows that $\rho(0, \tau)$ is the probability of finding the particle in the targeted state 0 at the time τ. Indeed because of the very low energy measured in the protocol thermal fluctuations play a role and the particle can be found in the wrong well at time τ, i.e. the proportion of success P_S of the procedure is equal to $\rho(0, \tau)$. In contrast the equilibrium distribution is $\rho^{eq}(0, \lambda(\tau)) = 1/2$. Then:

$$\left\langle e^{-\beta W(\tau)} \right\rangle_{\to 0} = \frac{1/2}{P_S} \tag{8}$$

Similarly for the trajectories that end the procedure in the wrong well (i.e. state 1) we have:

$$\left\langle e^{-\beta W(\tau)} \right\rangle_{\to 1} = \frac{1/2}{1 - P_S} \tag{9}$$

Taking into account the Jensen's inequality, i.e. $\left\langle e^{-x} \right\rangle \geq e^{-\langle x \rangle}$, we find that Eqs. (8) and (9) imply:

$$\begin{aligned} \langle W \rangle_{\to 0} &\geq k_B T \left[\ln(2) + \ln(P_S) \right] \\ \langle W \rangle_{\to 1} &\geq k_B T \left[\ln(2) + \ln(1 - P_S) \right] \end{aligned} \tag{10}$$

Notice that the mean work dissipated to realize the procedure is simply:

$$\langle W \rangle = P_S \times \langle W \rangle_{\to 0} + (1 - P_S) \times \langle W \rangle_{\to 1} \tag{11}$$

where $\langle . \rangle$ is the mean on all trajectories. Then using the previous inequalities it follows:

$$\langle W \rangle \geq k_B T \left[\ln(2) + P_S \ln(P_S) + (1 - P_S) \ln(1 - P_S) \right] \tag{12}$$

which is indeed the generalization of the Landauer's limit for $P_S < 1$. In the limit case where $P_S \to 1$, we have:

$$\left\langle e^{-\beta W} \right\rangle_{\to 0} = 1/2 \tag{13}$$

Since this result remains approximatively verified for proportions of success close enough to 100%, it explains why in the experiment we find $\Delta F_{\text{eff}} \approx k_B T \ln(2)$.

This result is useful because it strongly binds the generalized Jarzynski equality (a thermodynamic relation) to Landuer's bound.

References

1. R. Landauer, Information is physical. Phys. Today **44**(5), 23 (1991)
2. W. Weaver, C.E. Shannon, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 2010)
3. R. Clausius, *The Mechanical Theory of Heat* (McMillan, London, 1879)
4. E. Lutz, S. Ciliberto, Information: from Maxwell's demon to Landauer's eraser. Phys. Today **68**(9), 30 (2015)
5. J.M.R. Parrondo, J.M. Horowitz, T. Sagawa, Thermodynamics of information. Nat. Phys. **11**, 131 (2015)
6. H.S. Leff, A.F. Rex (eds.), *Maxwell's Demon: Entropy, Classical and Quantum Information, Computing* (Institute of Physics, Philadelphia, 2003)
7. M.B. Plenio, V. Vitelli, The physics of forgetting: Landauer's erasure principle and information theory. Contemp. Phys. **42**, 25 (2010)
8. K.K. Maruyama, F. Nori, V. Vedral, The physics of Maxwell's demon and information. Rev. Mod. Phys. **81**, 1 (2009)
9. L. Szilard, On the minimization of entropy in a thermodynamic system with interferences of intelligent beings. Z. Phys. **53**, 840 (1929)
10. C.H. Bennett, The thermodynamcis of computation-a review. Int. J. Theor. Phys. **21**, 905 (1982)
11. H.S. Leff, A.F. Rex (eds.), *Foundations of Statistical Mechanics* (Pergamon, Oxford, 1970)
12. L. Brillouin, *Science and Information Theory* (Academic Press, Waltham, 1956)
13. M.G. Raizen, Comprehensive control of atomic motion. *Science* **324**, 1403 (2009)
14. G.N. Price, S.T. Bannerman, K. Viering, E. Narevicius, Single-photon atomic cooling. Phys. Rev. Lett. **100**, 093004 (2008)
15. J.J. Thorn, E.A. Schoene, T. Li, D.A. Steck, Experimental realization of an optical one-way barrier for neutral atoms. Phys. Rev. Lett. **100**, 240407 (2008)

16. S. Toyabe, T. Sagawa, M. Ueda, M. Muneyuki, M. Sano, Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. Nat. Phys. **6**, 988–992 (2010)
17. T. Sagawa, M. Ueda, Generalized Jarzynski equality under nonequilibrium feedback control. Phys. Rev. Lett. **104**, 090602 (2010)
18. T. Sagawa, M. Ueda, Minimum energy cost for thermodynamic information processing: measurement and information erasure. Phys. Rev. Lett. **102**, 250602 (2009)
19. J.V. Koski, V.F. Maisi, T. Sagawa, J.P. Pekola, Experimental observation of the role of mutual information in the nonequilibrium dynamics of a Maxwell demon. Phys. Rev. Lett. **113**, 030601 (2014)
20. J.V. Koski, A. Kutvonen, I.M. Khaymovich, T. Ala-Nissila, J.P. Pekola, On-chip Maxwell's demon as an information-powered refrigerator. Phys. Rev. Lett. **115**, 260602 (2015)
21. J.V. Koski, J.P. Pekola, Maxwell's demons realized in electronic circuits. C.R. Phys. **17**, 1130–1138 (2016)
22. J.M. Horowitz, M. Esposito, Thermodynamics with continuous information flow. Phys. Rev. X **4**, 031015 (2014)
23. A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, E. Lutz, Experimental verification of Landauer's principle linking information and thermodynamics. Nature **483**, 187 (2012)
24. A. Bérut, A. Petrosyan, S. Ciliberto, Detailed Jarzynski equality applied to a logically irreversible procedure. Europhys. Lett. **103**(6), 60002 (2013)
25. M. Gavrilov, J. Bechhoefer, Arbitrarily slow, non-quasistatic, isothermal transformations. Europhys. Lett. **114**, 50002 (2016)
26. E. Aurell, K. Gawedzki, C. Meja-Monasterio, R. Mohayaee, P. Muratore-Ginanneschi, Refined second law of thermodynamics for fast random processes. J. Stat. Phys. **147**, 487 (2012)
27. S. Vaikuntanathan, C. Jarzynski, Dissipation and lag in irreversible processes. Europhys. Lett. **87**, 60005 (2009)
28. A.O. Orlov, C.S. Lent, C.C. Thorpe, G.P. Boechler, G.L. Snider, Experimental test of Landauer's principle at the sub- kb t level. Jpn. J. Appl. Phys. **51**(6S), 06FE10 (2012)
29. J.V. Koski, V.F. Maisi, J.P. Pekola, D.V. Averin, Experimental realization of a Szilard engine with a single electron. Proc. Natl. Acad. Sci. **111**, 13786 (2014)
30. Y. Jun, M. Gavrilov, J. Bechhoefer, High-precision test of Landauer's principle in a feedback trap. Phys. Rev. Lett. 113, 190601 (2014)
31. M. Gavrilov, J. Bechhoefer, Erasure without work in an asymmetric, double-well potential. Phys. Rev. Lett. **117**, 200601 (2016)
32. E. Roldàn, I.A. Martinez, J.M.R. Parrondo, D. Petrov, Universal features in the energetics of symmetry breaking. Nat. Phys. **10**, 457 (2014)
33. L. Martini, M. Pancaldi, M. Madami, P. Vavassori, G. Gubbiotti, S. Tacchi, F. Hartmann, M. Emmerling, S. Höfling, L. Worschech, G. Carlotti, Experimental and theoretical analysis of Landauer erasure in nano-magnetic switches of different sizes. Nano Energy **19**(Supplement C), 108–116 (2016)
34. J. Hong, B. Lambson, S. Dhuey, J. Bokor, Experimental test of Landauer's principle in single-bit operations on nanomagnetic memory bits. Sci. Adv. **2**, e1501492 (2016)
35. P.A. Camati, J.P.S. Peterson, T.B. Batalhão, K. Micadei, A.M. Souza, R.S. Sarthour, I.S. Oliveira, R.M. Serra. Experimental rectification of entropy production by Maxwell's demon in a quantum system. Phys. Rev. Lett. **117**, 240502 (2016)
36. N. Cottet, S. Jezouin, L. Bretheau, P. Campagne-Ibarcq, Q. Ficheux, J. Anders, A. Auffèves, R. Azouit, P. Rouchon, B. Huard. Observing a quantum Maxwell demon at work. Proc. Natl. Acad. Sci. U. S. A. **114**(29), 7561–7564 (2017)
37. M.A. Ciampini, L. Mancino, A. Orieux, C. Vigliar, P. Mataloni, M. Paternostro, M. Barbieri, Experimental extractable work-based multipartite separability criteria. npj Quantum Inf. **3**(1), 10 (2017)
38. S.M. Herre, S.J. van der Zant, F. Luis, R. Gaudenzi, E. Burzurí, Quantum-enhanced Landauer erasure and storage. arXiv:1703.04607

39. E. Lubkin, Keeping the entropy of measurement: Szilard revisited. Int. J. Theor. Phys. **26**(6), 523–535 (1987). ISSN 1572–9575
40. V. Vedral, Landauer's erasure, error correction and entanglement. Proc. R. Soc. Lond. A: Math. Phys. Eng. Sci. **456**(1996), 969–984 (2000)
41. M.P. Frank, The physical limits of computing. Comput. Sci. Eng. **4**, 16 (2002)
42. E. Pop, Energy dissipation and transport in nanoscale devices. Nano Res. **3**, 147 (2010)
43. K. Chida, S. Desai, K. Nishiguchi, A. Fujiwara, Power generator driven by Maxwell's demon. Nat. Commun. **8**, 15310 (2017)
44. D. Bray, Energetic costs of cellular computation. Nature **376**, 307 (1995)
45. P. Mehta, D.J. Schwab, Energetic costs of cellular computation. Proc. Natl. Acad. Sci. **109**, 17978 (2012)
46. Y. Tu, The nonequilibrium mechanism for ultrasensitivity in a biological switch: sensing by Maxwell's demons. Proc. Natl. Acad. Sci. **105**, 11737 (2008)
47. A.C. Barato, D. Hartich, U. Seifert, Efficiency of cellular information processing. New J. Phys. **16**, 103024 (2014)
48. S. Ito, T. Sagawa, Maxwell's demon in biochemical signal transduction with feedback loop. Nat. Comm. **6**, 7498 (2015)
49. A.C. Barato, U. Seifert, Cost and precision of Brownian clocks. Phys. Rev. X **6**, 041053 (2016)
50. T.E. Ouldridge, C.C. Govern, P. Rein ten Wolde, Thermodynamics of computational copying in biochemical systems. Phys. Rev. X **7**, 021004 (2017)
51. K. Sekimoto, *Stochastic Energetics*. *Lecture Notes in Physics*, vol. 799 (Springer, 2010)
52. U. Seifert, Stochastic thermodynamics, fluctuation theorems and molecular machines. Rep. Prog. Phys. **75**(12), 126001 (2012). http://stacks.iop.org/0034-4885/75/i=12/a=126001
53. S. Ciliberto, Experiments in stochastic thermodynamics: short history and perspectives. Phys. Rev. X **7**, 021051 (2017)
54. C. Jarzynski, Nonequilibrium equality for free energy differences. Phys. Rev. Lett. **78**, 2690 (1997)
55. C. Jarzynski, Equilibrium free-energy differences from nonequilibrium measurements: a master-equation approach. Phys. Rev. E **56**, 5018 (1997)

Experimental Tests of the Landauer Principle in Electron Circuits, and Quasi-Adiabatic Computing Systems

Alexei O. Orlov, Ismo K. Hänninen, César O. Campos-Aguillón, Rene Celis-Cordova, Michael S. McConnell, Gergo P. Szakmany, Cameron C. Thorpe, Brian T. Appleton, Graham P. Boechler, Craig S. Lent, and Gregory L. Snider

Contents

A. O. Orlov (✉) · I. K. Hänninen · C. O. Campos-Aguillón · R. Celis-Cordova
M. S. McConnell · G. P. Szakmany · C. C. Thorpe · B. T. Appleton · G. P. Boechler
C. S. Lent · G. L. Snider
Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA

C. S. Lent et al. (eds.), *Energy Limits in Computation*,
https://doi.org/10.1007/978-3-319-93458-7_6

1 Energy in Computation

1.1 Introduction

The development of integrated circuits is limited by power dissipation. Anyone using a mobile device has experienced the effects of the high dissipation required in today's computing devices: laptop computers that can burn the lap, and hand-held devices with a short battery life. The CPU is perhaps the most important component in these systems since it processes the information, and this information processing requires energy. The CPU dissipates a significant fraction of the energy used in computational systems, so improvements in CPU energy efficiency would pay large dividends.

Is it possible to do computation with less dissipation than with standard CMOS? Is there a lower limit to the energy that must be dissipated to heat in the processing of information? Recently, there have been a number of papers, for example [1], that suggest that charge-based computation faces fundamental energy dissipation limits, usually stated as k_BT ln2 per bit operation, spurring an enormous research effort searching for alternative state variables to encode information. The energy k_BT ln2 is often called the "Landauer Limit," though this is a misnomer. It should properly be called the "Ultimate Shannon Limit" (USL) [2], the energy needed to maintain a signal-to-noise ratio of unity in a thermal environment. The "Landauer Principle" (LP) states that an energy of k_BT ln2 is necessarily dissipated to heat only when information is destroyed, or "erased" and that no lower limit exists if information is preserved. Thus, there is no Landauer Limit. If the Landauer Principle is correct, energy savings are possible by avoiding the destruction of information, using reversible computing with adiabatic logic transitions.

The Landauer Principle is widely accepted, but there is a significant literature that either dismisses it outright or asserts that theoretical proofs put forward are flawed [3, 4]. Experimental tests of LP were not performed because the energy involved, k_BT ln2 $\sim$ 3 zJ at room temperature, was considered immeasurably small. The reduction in power dissipation in adiabatic circuits was measured by thermoelectric techniques [5], but not at a level that could resolve energies of a few k_BT. Recently an experiment by Berut et al. [6] verified one half of the Landauer Principle, that a minimum energy of k_BT ln2, the Ultimate Shannon Limit, must be dissipated in an irreversible operation where information is destroyed. We have recently made experimental measurements that test the other assertion in the Landauer Principle, that there is no lower limit when information is not destroyed [7].

The conventional logic used in today's computers encodes information with charge stored on capacitors, specifically the CMOS transistor gates and interconnect capacitance. Power dissipation for standard CMOS logic is given by the equation

$$P_{\text{Total}} = N\left(\alpha C V_{\text{DD}}^2 f + P_{\text{passive}}\right) \tag{1}$$

where V_{DD} is the supply voltage, C is the load capacitance at the output of each logic gate, N is the number of gates, α is the activity factor (the fraction of

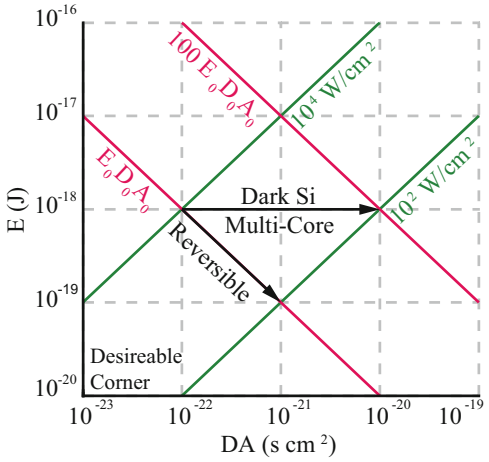

Fig. 1 EDA Comparison, where green lines are lines of constant power density, and the red line represents constant resource utilization

devices switching), and f is the operating frequency. The first term represents the active power dissipation, which is the power dissipated in information processing. The second term, the passive power dissipation, is power that is simply wasted when a voltage is applied to the circuit, due to transistor leakage. Reduction of the passive power has been the subject of intense research, such as that in the areas of tunnel FETs and ferro-electric gate FETs. However, even if the passive dissipation can be eliminated, the dissipation in information processing can increase to unacceptable levels. Alternative approaches such as neuromorphic computing and quantum computing can reduce energy use, but are limited in applicability, accuracy or both. An approach is needed that can provide significant energy savings while giving deterministic "correct" answers.

Equation (1) and Fig. 1 highlight the twin problems faced by the CMOS electronics industry. The ITRS 2015 Roadmap projects a fully scaled CMOS process to have a device density of 10^{10} cm^{-2}, a switching speed of 12 THz, and a switching energy of 3 aJ (750 k_BT at room temperature). As an extreme example of possible energy scales, if all of the devices on such a chip were switched at full speed, just the *active* power dissipation of the chip would be approximately 150 kW/cm^2. Even lowering the switching energy to 100 k_BT, a practical limit for deterministic computing with negligible error rates, will only reduce the active power to 20 kW/cm^2. While this is an extreme example, clearly the processing of information using current methods does not provide a path to ultra-high-density high-speed computation where all devices are switched at their maximum operating frequency. When 3D circuits are considered, the power density challenge becomes even more daunting, with current 3D schemes limited to a single logic layer, and other layers restricted to low-activity circuits such as memory. Whatever the circuit and device implementations, cooling imposes limits on power density, and this amounts to limits on the computational density.

While the industry seeks to deal with the problem of dissipation with approaches such as steep devices and dark silicon [8–11], it is important to understand the fundamental limits of dissipation in computation. Discussions of the minimum amount of energy that must be lost to heat during computation can be tied to debates going back to Maxwell's demon. In 1961 Landauer [12] postulated that energy must be dissipated as heat only when information is destroyed, an idea that has come to be known as the Landauer principle (LP). The minimum amount of energy that must be dissipated is related to a quantity known as the Ultimate Shannon Limit [2, 13], k_BT ln2, the minimum energy needed to create a bit of information that is distinguishable from noise. According to LP, if information is not destroyed there is no fundamental lower limit to dissipation in computation, only practical limits due to system requirements that can be much, much less than k_BT.

The Landauer principle is a consequence of how the energy used to represent information interacts with the environment, and applies to all state variables [14]. Approaches such as neuromorphic and analog computing do not circumvent LP, but trade accuracy for power dissipation, much as our brain does. While this approach is adequate for some classes of problems such as image processing, at which our brain excels, solutions to other problems require accurate answers. Only reversible computing can break through the Shannon limit by recovering and reusing the bit energy.

1.2 A New Look at Reversible Computing and Energy Recovery

The electronics industry is locked into the dissipation limitations discussed above by standard CMOS circuitry, where the dissipated active power described in Eq. (1) is a consequence of the fact that the information contained in the logic gate is destroyed at every logic transition. The energy to form a bit of information is drawn from the power supply, and when it is no longer needed it is discarded to heat. This has the advantage of being very simple, as decisions to destroy bits of information are made locally. Discarding information as heat is very convenient, as dealing with destroyed bits is not a device or circuit problem, but someone else's problem, namely the thermal manager's.

In each logic gate, each bit of information is represented by the energy

$$E_{\text{Bit}} = \frac{1}{2} C V_{\text{DD}}^2 \tag{2}$$

stored on the capacitor C, and this entire amount of energy is dissipated as heat twice in each cycle as the bit of information is created and then destroyed. Standard CMOS circuits make this destruction of information unavoidable, so the only way to limit active dissipation is to reduce the energy in a bit (reduce V_{DD}) or limit the rate at which bits are destroyed (limit f).

There is another approach. Reversible computing with adiabatic clocking can reduce the active power and break the connection between active and passive dissipation. The energy dissipation with adiabatic clocking is, assuming a constant capacitance and $T_{\text{ramp}} >> RC$:

$$E_{\text{Adiabatic}} = RC^2 V_{\text{DD}}^2 \Big/ T_{\text{ramp}} \tag{3}$$

where R is the resistance associated with the circuit and T_{ramp} is the duration of the clock circuit ramp. When the ramp time is greater than the RC time constant of the circuit, this approach offers energy efficiencies orders of magnitude better than current computational paradigms. Reversible adiabatic computing is an idea that was proposed many years ago, but was dismissed as "slow," since the circuit was not run "as fast as possible,", as well as assertions that reversible systems simply cannot reduce dissipation, because the Landauer principle itself is incorrect. However, with the freeze in clock speeds since the early 2000s, circuits already operate well below speeds set by RC limits, so a trade-off of clock speed for dissipation is already being made. Since speeds set by the RC limit are so high, reversible designs can be quite fast. Circuits once considered too "slow" can therefore become attractive, and experiments have shown that significant power savings are possible.

In reversible systems, dissipation is ultimately limited by leakage. While approaches such as steep devices can reduce leakage, more is likely needed for reversible adiabatic systems to become practical. New devices are needed that eliminate leakage currents, along with a reversible architecture that enables energy recovery.

1.3 When Does Energy Recovery Make Sense?

An important question in system design is: How fast should a computational chip be run. The answer, of course, is that it should run "as fast as possible"! But how fast is this? In the 1980s and 1990s the answer was simple: The chip should run as fast at the RC time constants set by devices and circuits would allow, which means "as fast as physically possible." Reversible, adiabatic circuits were examined during this time, but were not pursued because they ran slower than the limit imposed by the circuit, which made little sense at the time since dissipation was not a problem. Since circuit speed was the only important performance criterion, with power dissipation being of little concern, reversible adiabatic circuitry was abandoned in favor of pure device scaling. However, by the late 1990s the resulting exponential increase in the power density made further scaling with "as fast as physically possible" architectures untenable.

Dark silicon, large on-chip caches, and multi-core arose as techniques to run "as fast as possible," but at a power density limited by cooling. In dark silicon,

power is removed from logic gates that are not being used. This eliminates the leakage power from the gates, but more importantly reduces the overall power density since the heat generated in the active areas is spread over the dark areas. Hence, dark silicon effectively reduces device density. At the 8 nm node it has been estimated that up to 80% of the chip will be dark at any given time [10]. The multi-core approach uses a lower-than-possible clock frequency (increasing the effective delay) to limit dissipation, and recovers some of the lost performance through parallelism. While giving up raw performance, at least these approaches provide some computational performance improvements by making use of the ever-increasing number of transistors that process developments give, in the form of Moore's law. They do not change the basic approach of energy use in CMOS, an approach that can uncharitably be characterized as drunken sailor mode. As money flows through the fingers of a drunken sailor, so energy flows and is wasted in CMOS. While this is a simple approach, it is very wasteful. There may be gains to be had by examining the problem at a basic level.

Dark silicon and multi-core are ways to keep the processor running "as fast as possible." The real question is, what does "as fast as possible" REALLY mean? Since "as fast as physically possible" is no longer possible, which approach is best? The goal in computation is to make the best use of limited resources. Energy, time (delay), and area can be viewed as resources available to the computation, so an appropriate figure of merit for the best technology is one with the smallest product of energy, delay, and area, EDA. While these are the traditional resources considered for computation, there are other resources whose limits must be considered. Cooling is a resource whose limit of 100–200 W/cm^2 for air-cooling now poses the greatest challenge for electronics, and should be considered in evaluating technology choices. For computation in a mobile platform the energy source, the battery, is a limited resource, and a variant of the EDA metric is useful. In this case the goal is to minimize the energy drawn from the battery, so an appropriate metric is E^2DA.

The EDA metric is a useful tool in evaluating technology choices in the face of limited resources. A given technology will yield some intrinsic performance for a basic logic element, the switching energy E_o, the propagation delay D_o, and the area A_o. Using the EDA metric, the best performance possible for this logic is $E_oD_oA_o$. If there are no other constraints, the best technology is that with the smallest $E_oD_oA_o$, and it makes no sense to trade performance for energy savings using reversible computing. However, if cooling is a limited resource, it is necessary to re-evaluate the use of the resources E, D, and A. To meet the power density limit, performance must be given up. But what is the best way to trade away performance? For example, consider a circuit with intrinsic performance parameters of $E_o = 1$ aJ, $D_o = 10$ ps, and $A_o = 10^{-11}$ cm^2, shown in Fig. 1 where the delay and area have been combined to produce a 2D plot. At full performance, this circuit would have a power density of 1×10^4 W/cm^2, so the dissipation must be reduced by a factor of 100 to meet the cooling constraint. Using the multi-core approach the delay must increase by a factor of 100 (compared to what is possible without the power constraint), so the EDA metric becomes 100 E_o D_o A_o. Likewise, if the dark silicon approach is taken, the effective area of the active circuit is increased

by a factor of 100, so the EDA again becomes 100 $E_o\ D_o\ A_o$. This means that the circuit is not making the best use of the available energy, time, and area, and some other approach might be better. In adiabatic reversible logic, performance is traded in time, but as a result the switching energy decreases. When the delay is increased by a factor of 10 in adiabatic reversible logic the dissipated energy is reduced by a factor of 10 and the EDA metric remains $E_o\ D_o\ A_o$, since the factors of 10 cancel, while the power density E/(DA) is lowered by a factor of 100 as required. This demonstrates that when the computation is resource constrained, in this case by cooling, adiabatic reversible computing makes sense because it makes the best use of available resources. In a mobile computing environment, adiabatic reversible computing is even more attractive since the energy saved (recycled) is squared in the metric E^2DA, reflecting the premium placed on battery life. Here, adiabatic reversible computing will be most attractive in computationally intensive applications where simply turning the device off most of the time is not acceptable.

Adiabatic reversible computing is an alternative way to compromise computing performance to meet the constraints of limited resources. However, to realize any of the advantages that reversible computing offers, the bit energies must be recovered and recycled.

2 Experimental Test of the Landauer Principle

2.1 Experimental Overview

As a fundamental test of the Landauer principle, we will demonstrate experimentally that energy dissipation to heat in a *reversible* computational scheme can be set arbitrarily low while the bit energy is maintained well above k_BT, by appropriately choosing the degree of adiabaticity. We will show that this very small dissipated energy, about 20 yJ, can be accurately measured in the presence of thermal noise. Next, we will show that the energy dissipation in the *irreversible* bit manipulation inevitably leads to dissipation of the entire bit energy, and since minimum bit energy is k_BT ln2 (equivalent to USL), this sets a lower bound on the energy that must be dissipated in an irreversible process.

For the experimental test of LP we chose a simple analog of a bipolar clock circuit described in the section of this chapter on adiabatic circuits. In such a system, the binary "zero" and "one" are represented by the respective positive ($+V_0$) and negative voltage ($-V_0$) values across a capacitor, so that the voltage that represents a bit is $V_C = \pm V_0$. A discharged capacitor represents a "NULL" state which does not carry any information, $V_{\mathrm{NULL}} \approx 0$. Note that the system is symmetric in that 0 and 1 are each represented by the same amount of energy. The functional diagram of the circuit used in the experiment is shown schematically in Fig. 2a. The first step is to copy a bit of information into the memory cell M (in our case a capacitor accessed through a resistor, and a resistor is representing a channel of a transistor in a CMOS

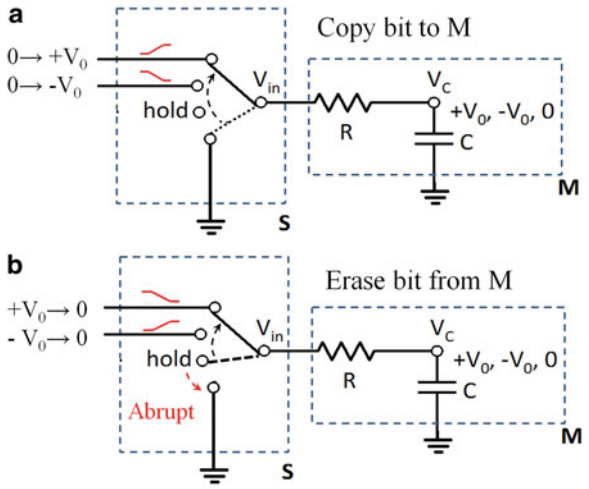

Fig. 2 (**a**) Functional diagram of copy operation. (**b**) Functional diagram of erase operation

logic circuit). Schematically a switch is used to select whether a 1, 0 or NULL is written into the cell, which is initially at NULL. A 1 or 0 is written into the cell by quasi-adiabatically ramping the input voltage from 0 V to $+V_0$ or $-V_0$. When the ramp reaches its final value, the switch can be placed in the hold position, and a bit of information is then held in the capacitor, represented as charge with an energy of

$$E_{\mathrm{BIT}} = C\left(\frac{V_0^2}{2}\right) \tag{4}$$

The bit energy, E_{BIT}, must be greater than the value defined by USL, $k_B T$ ln2 to make the information distinguishable from noise [2]. Erasing the bit of information stored in M can be done in one of three ways, as shown schematically in Fig. 2b, where the voltages on the three input lines are at $+V_0$, $-V_0$, and 0 V. Once the switch has been moved to the desired position, the voltages $+V_0$ and $-V_0$ are adiabatically ramped to zero. A decision must be made as to which switch position to choose when erasing the bit of information in M. If one knows the value stored in M, then the switch can be set to the appropriate position, $+V_0$ or $-V_0$ and the voltage ramped down with little dissipation. However, this knowledge of the contents of M means that a copy of the information exists somewhere. If the information exists only in M, then no choice can be made that requires knowledge of the information in M. In this case one either can connect the switch to the ground discharging the capacitor non-adiabatically through the resistor ("abrupt erase") or make a guess at the information contained in M with a ½ probability of choosing the wrong switch position (e.g., switching from $+V$ to $-V$). If the capacitor is discharged abruptly by connecting it to ground, the full energy of the bit, E_{BIT}, is dissipated to heat in the resistor R.

Guessing the state of the memory cell incorrectly results in a dissipation of 4 E_{BIT} since the voltage across the resistor is abruptly doubled. This experiment, then, is a test of the Landauer principle, in that $E_{\mathrm{Bit}} > k_B T$ ln2 must be dissipated only if information is destroyed. If a copy exists, information is not destroyed in the erasure and dissipation can be less than $k_B T$ ln2.

2.2 Experimental and Measurement Details

Taking into consideration the limits for deterministic computing mentioned previously, we settled on a bit energy value of $E_{\mathrm{BIT}} = 72\ k_B T$, which corresponds to a probability of "thermally activated error[1]", $p_{\mathrm{therm}} \sim exp(-E_{\mathrm{BIT}}/k_B T) \sim 5 \times 10^{-32}$, where $k_B T = 4.05 \times 10^{-21}$ *J at 293 K.* This probability guarantees the avoidance of thermal errors in 30 years ($\sim 10^9$ s) of operations, assuming clock speed of 100 GHz, and 10^{11} devices on chip.

The circuit elements, R and C, must be selected to allow measurements of electrical signals (voltages) corresponding to very small energies ($<<k_B T$). The voltage across the capacitor for a given E_{BIT}, in our case chosen to be 75 μV,

$$|V_0| = \sqrt{\frac{2E_{\mathrm{BIT}}}{C}} \approx 75\ \mu\mathrm{V}, \tag{5}$$

is determined by a capacitance value, and larger voltage can be chosen by reducing the value of the capacitor. The capacitor C for the experiment in Fig. 2 is chosen to represent the largest capacitor in the circuit so that parasitic capacitances (dominated by the input capacitance of the measurement amplifier, $C_{\mathrm{IN}} \approx 3$ pF) do not contribute a significant error to the measurements. We chose a D series low loss ceramic disk capacitor with rated insulation resistance 10^{10} Ω at 100 V. The value of the capacitor, 104.4 pF, was measured with capacitance bridge. To reduce stray capacitances of the cables the capacitor C is soldered directly to the same PCB pad as the input terminal of the first stage amplifier.

To minimize energy losses in the resistor R one needs to use linear ramps for transitions between logic levels, thus achieving quasi-adiabatic charging of capacitors. The voltage that develops across R during linear voltage ramps is inversely proportional to the ramp time, t_1:

$$V_{\mathrm{R\_sweep}} = RC\frac{dV_{\mathrm{IN}}}{dt} = RC\left(\frac{V_0}{t_1}\right) \tag{6}$$

[1]The above estimate gives a probability of the bit value randomly acquiring an incorrect value due to thermal fluctuations of voltage across the capacitor and it assumes *an ideal noiseless readout circuit*.

where V_0 is the final (initial) voltage for the ramp up (down). The energy dissipated in the resistor R during the ramp process is:

$$E_R = \frac{{V_R}^2}{R} t_1 = V_0^2 C \left(\frac{RC}{t_1} \right) = 2E_{\mathrm{BIT}} \left(\frac{RC}{t_1} \right) \tag{7}$$

which can be made as small as desired for a given set of circuit parameters (V_0, R, C) by the appropriate choice of t_1. For the purpose of demonstration, we choose the amount of energy dissipated as heat for each ramp $E_R = 0.005\ k_B T$. There is no strict limitation on the choice of the resistor R value; we chose it based on the following considerations. First, the resistor value must be relatively small so that we can neglect all parasitic impedances parallel to it in the bandwidth (BW) of interest (see below). Second, it can be seen from (6) and (7) that the ratio RC/t_1 is a constant for a given E_R and V_0. Based upon the first constraint (negligible parasitics) we have chosen for R a 1.1 k Ω metal film chip resistor. This sets the ramp time, using (7), at $t_1 = 3.36$ ms. As we show below, to extract the signal from noise we have to resort to multiple curve averaging, and there is a "reasonable experimental" upper limit on a ramp time t_1, that comes from the averaging time required to perform experiments with large number of averaged signal traces. The above choice of t_1 allows us to repeat the entire operation cycle ($T \approx 18$ ms) up to four million times within a 20-h period. The chosen time constant of the RC chain,[2] $\tau = RC \approx 0.11$ μs, insures that voltage across the capacitor reaches 99.9% of the input voltage step within ~0.8 μs, a negligibly short time compared to ramp time t_1. The voltage across resistor during the ramp time is given by:

$$V_R = \sqrt{\frac{E_R R}{t_1}} \approx 2.5\ \mathrm{nV}. \tag{8}$$

This signal is challenging to measure in the presence of noise, which is many orders of magnitude greater than V_R. Indeed, at any given temperature, thermal agitation of the electron cloud leads to voltage fluctuations across the capacitor in any RC circuit, with the RMS value of[3]

$$V_N = \sqrt{\frac{4k_B T R}{4RC}} = \sqrt{\frac{k_B T}{C}}. \tag{9}$$

Note that this value of noise voltage corresponds to the energy of $E = k_B T/2$ stored in the capacitor. This also can be interpreted as the fact that any capacitor has one thermodynamic degree of freedom, so that the mean energy stored in it

[2]The parameter τ also defines the response time for the acquisition system to be able to capture the power dissipation process in the ERASE WITHOUT A COPY experiment that corresponds to a discharge of the capacitor C through resistor R.

[3]This formula takes into account the "brickwall" bandwidth of an RC circuit, $B = \frac{\pi}{2} \frac{1}{2\pi RC}$.

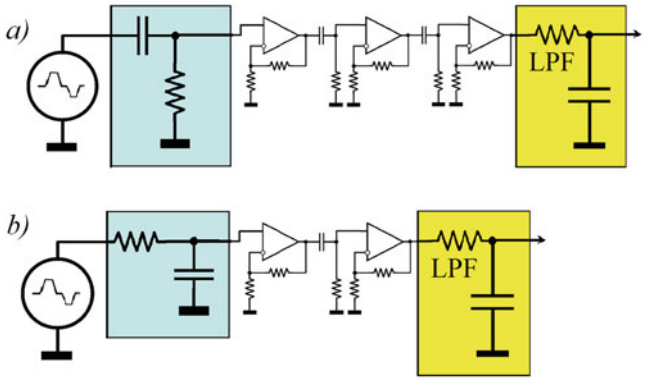

Fig. 3 Simplified circuit diagrams of the measurement setup to measure (**a**) V_C and (**b**) V_R. Elements in the blue boxes are the components R and C referred to in the text. Elements in the yellow box form a low pass filter with $f_H = 14$ kHz. A Tektronix AFG 3252 arbitrary waveform generator with a two-stage, low output resistance (10 Ω) voltage divider is used as a voltage source. The high-pass filter between stages necessary to prevent overload of the subsequent stages due to random DC offset in the first and second stages is set to 0.16 Hz

at a given temperature is $k_BT/2$ [15]. Given our choice of parameters, the thermal voltage across R in a full bandwidth of $1/4\,RC = 2.185$ MHz is $V_N = 6.24$ μV. This estimate shows that $V_N/V_R > 2000$ if the full bandwidth is explored. However, the measurement bandwidth (MBW) needs to be just broad enough to avoid the loss of the signal during the transient time, $f_T \geq 30/t_1 \approx 10$ kHz. In this experiment, we choose the MBW ≈ 14 kHz by using a 1st order low-pass filter at the output of the amplifier, which corresponds to the equivalent brick-wall noise bandwidth of ≈20 kHz.

The measurement system used to measure this voltage is comprised of a 3-stage amplifier with a total gain of 79.3 × 79.3 × 79.3 ≈ 500,000, Fig. 3a. The referred-to-input RMS total noise voltage in the measurement system can be calculated:

$$V_N = \sqrt{B\left(e_N^2 + i_N^2 R^2 + 4k_B T R\right)} \tag{10}$$

where e_N and i_N are voltage and current noise spectral densities (NSD) of the amplifier, and $4k_BTR$ is the Johnson NSD in the resistor R. The noise in the measurement system is, in turn, dominated by the first stage JFET input amplifier (TLE2081) with voltage NSD,[4] $e_N = 11.6$ nV/Hz$^{1/2}$, and the Johnson NSD in the resistor R (4.2 nV/Hz$^{1/2}$). The noise voltage developed across R due to current NSD of the amplifier (2.8 fA/Hz$^{1/2}$) is negligibly small compared to the first two

[4]The noise contributions of the second and third stage in the amplifier are negligibly small due to high gain (79.3) of the first stage amplifier.

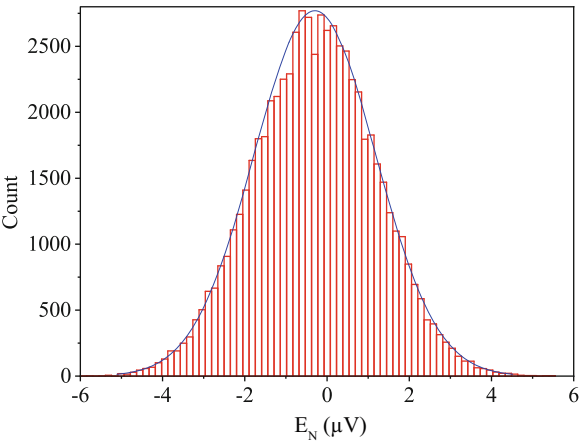

Fig. 4 The measurement of referred to input measurement amplifier noise (first stage is op-amp TLE2081). The size of one bin corresponds to 1 bit of ADC resolution. Solid line—Gaussian with a standard deviation of 1.48 μV. The value calculated from datasheet for the noise bandwidth of 20 kHz is 1.7 μV

terms. Equation (10) gives, for the resulting total noise in the 20 kHz bandwidth, $V_N \approx 1.7$ μV. To characterize the noise of the amplifier experimentally, we acquire signal traces at the output of the amplifier with no signal applied to the input. Using an ADC with acquisition rate 200 ksps,[5] 100 ms time frames are acquired and then statistically analyzed to calculate the standard deviation, V_N. Figure 4 shows a histogram of an acquisition snapshot containing 2^{16} samples (referred to input). As expected, the extracted RMS value of voltage noise at the amplifier input, $V_N \approx 1.5$ μV, is very close to that expected from Eq. (10). Obviously, the signal we are trying to measure, $V_R \sim 2.5$ nV, will still be completely obscured by this noise. However, a simple solution, trace averaging, exists for enhancing the SNR in the case where the noise is stationary Gaussian noise, as in Fig. 4. By averaging enough traces, N, one can achieve any desired noise reduction. However, it is literally a matter of time, because measured noise voltage scales as $1/\sqrt{N}$, so enhancement in signal-to-noise ratio, SNR, defined here simply as SNR $= V_R/V_N$ is achieved at the expense of a squared acquisition time. It should be noted that signal averaging to reduce noise does not mean that actual dissipation is being ignored. The measured signal contains a contribution from the signal as well as thermal noise. Averaging reduces the thermal noise contribution but not the contribution of dissipation from the input signal. By averaging 4×10^6 traces over 20 h, we improved SNR by a factor of $\sqrt{\mathrm{N}} = 2000$, and reduced the noise down to a tolerable level of $V_N \approx 1.5$ μV/2000 $= 0.75$ nV. Note that the presence of white Gaussian noise at the digitizer's input improves the resolution by nearly 11 bits due to the oversampling/decimation mechanism and the quantization error can be neglected if the noise level is greater than least significant bit (LSB) of the ADC [16].

Unlike V_R, the voltage across the capacitor, $V_C \approx 75$ μV, can be readily measured for a bit represented by 72 k_BT. Indeed, even without averaging the ratio $V_0/V_N > 30$ makes the signal clearly distinguishable from noise. The total gain of the amplifier

[5] At this acquisition rate, we can capture the highest frequency components of the spectrum, 22 kHz.

for this measurement is also reduced (down to ~3500) since the signal is so much stronger and only two stages, Fig. 3b, are required to achieve levels appropriate for the digitizer.

2.3 Experimental Results and Discussion

The results of the experiment outlined in Sect. 2.2 are shown in Fig. 5. The top panel shows the input voltage applied to the memory cell M, where a "1" and then a "0" are copied into the cell and then erased without destroying the information

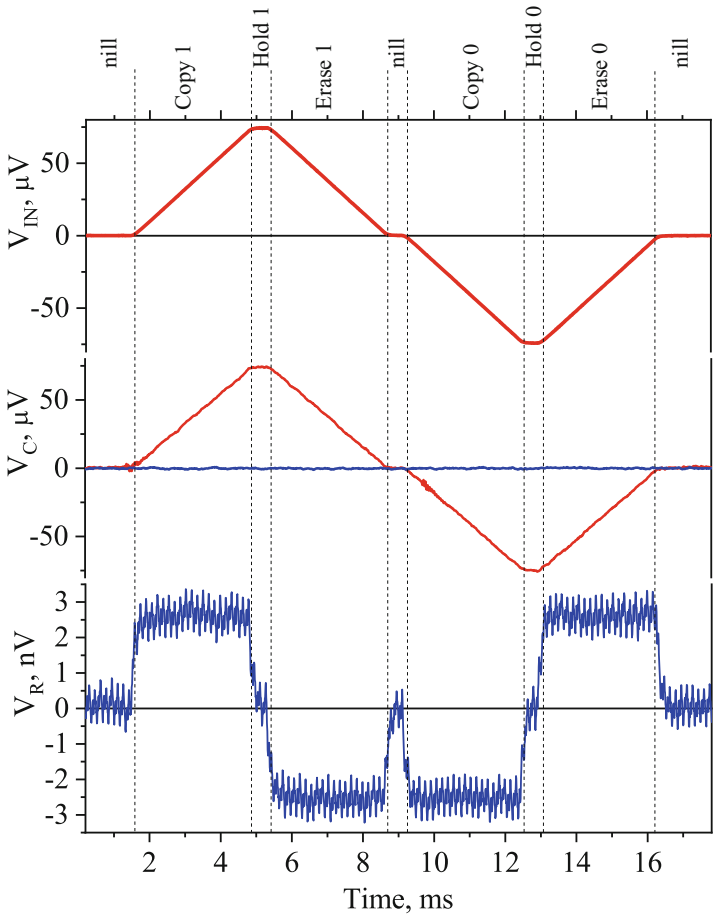

Fig. 5 Waveforms for COPY–ERASE WITH A COPY experiment. (**a**) Input waveform applied to the RC "Memory Cell." (**b**) Voltages across the resistor (blue) and capacitor (red) plotted on the same scale; (**c**) voltage across the resistor averaged 4×10^6 times. $R = 1.1$ kΩ, $C = 104$ pF; $T = 293$ K

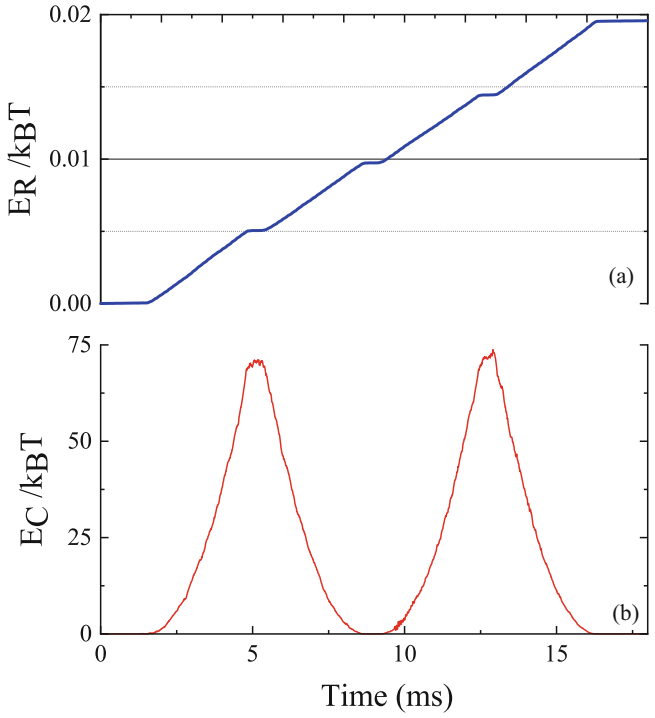

Fig. 6 Energy balance for the COPY–ERASE WITH A COPY experiment. (**a**) Energy dissipated in the resistor $R = 1.1$ kΩ (**b**) Energy delivered to the capacitor $C = 104$ pF; $T = 293$ K

(a copy is kept in the experimental system so we know how to erase with little dissipation). $V_0 = 75$ μV is chosen so that $E_{\text{Bit}} \approx 72\ k_BT$ of charge is delivered to the capacitor to represent the information. The middle panel shows a single-trace (no averaging) measurement of the voltage across the capacitor (red) along with voltage across the resistor measured with amplifier of Fig. 3b. This shows that a robust bit of information is delivered to the capacitor, with a very good signal-to-noise ratio, while the voltage across the resistor is so small that it is lost in the noise. As discussed, an accurate measurement of the energy lost in the resistor requires averaging to eliminate the noise signal. The bottom panel in Fig. 5 shows the result of voltage measurement across R obtained by averaging 4×10^6 traces. Approximately 0.005 k_BT, the amount equal to 20 yJ, of energy is lost to heat in the resistor during the charge and discharge (ramp) phases. This is far below k_BT, in accordance with the Landauer principle, and shows that a logic transition can be made with very little dissipation, as long as information is not destroyed. This experiment represents, to our knowledge, the lowest energy dissipation directly measured to date.

Figure 6 shows the calculation based upon experimental data Fig. 5 of energy dissipation for a full cycle of operations COPY 1–HOLD–ERASE WITH A

COPY TO NULL, COPY 0–HOLD–ERASE WITH A COPY TO NULL along with the energy delivered/removed from the capacitor. The total energy dissipated increases during each of the COPY and ERASE WITH A COPY operations, but the dissipation during each phase of the cycle is clearly much less than k_BT.

One can ask how much energy needs to be dissipated at the input of the measurement apparatus to perform a measurement? The measurement is itself a COPY operation that in principle can be performed with arbitrarily small dissipation. To estimate this energy loss in our experiment we need to consider a load on the energy storing capacitor C by the leakage resistance of the capacitor C ($>10^{10}$ Ω) and the input resistance of the amplifier ($>10^{12}$ Ω). A conservative estimate for the energy loss in the measurement circuit is on the order of 10^{-7} E_{BIT}, or $<10^{-4}$ $E_R \approx 10^{-6}$ k_BT.

Finally, let us consider a case where a bit of information was stored in our system, but then this information was irreversibly erased by closing the switch from "hold" to "ground." This situation corresponds to the ERASE WITHOUT A COPY experiment and as expected in such a case, all of the bit energy is dissipated to heat in the process. If the bit energy is the smallest discernable from noise, k_BT ln2, then all of it will be dissipated thus confirming the so-called "Landauer limit," really the ultimate Shannon limit. For experimental illustration, we have chosen $E_{\mathrm{bit}} = 30\, k_BT$, which corresponds to $V_0 = 50\ \mu\mathrm{V}$. In contrast to an adiabatic charging process, ERASE WITHOUT COPY is an abrupt process which happens in a much shorter timeframe, on the order of $\tau = RC$. Therefore, it requires a different (high speed) amplifier and acquisition system (broadband digital scope). A two-stage high speed amplifier (BW > 10 MHz) with a total gain of ~6300 employing a low noise videoamplifier AD811 was used for this purpose.[6] The task of the experiment is to capture an exponential decay for discharging capacitor, $V_R = V_0 \exp\left(-\frac{t}{RC}\right)$. The result of integration $E_{\mathrm{diss}} = \frac{1}{R}\int_0^T V_R^2(t)dt$ in this case should produce the known result, $E_{\mathrm{diss}} = \frac{CV_0^2}{2}$, i.e., the energy stored in the capacitor C is dissipated to heat when it is abruptly discharged through a resistor R. The results of the measurements are shown in Fig. 7. The exponential decay is observable, however, the front of the pulse ($t < 0.1\ \mu\mathrm{s}$), is distorted due to the limited bandwidth of the acquisition system so only about 60% of the area under pulse is captured. For comparison, we plotted a simulated exponential decay from $V_0 = 50\ \mu\mathrm{V}$ to zero (dashed red line in Fig. 7). The result of integration (inset in Fig. 7) verifies that measured energy dissipation in this process ≈21 k_BT is about 60% of the full energy stored in the capacitor, $E_{\mathrm{diss}} = \frac{CV_0^2}{2} = 30\, k_BT$.

[6]Once again, trace averaging is used to improve the SNR since much larger bandwidth and large current NSD of AD811 results in $V_N \sim 100\ \mu\mathrm{V}$. Note that it takes only 100 ms to average 10^5 traces and effectively reduce the noise to 0.3 μV.

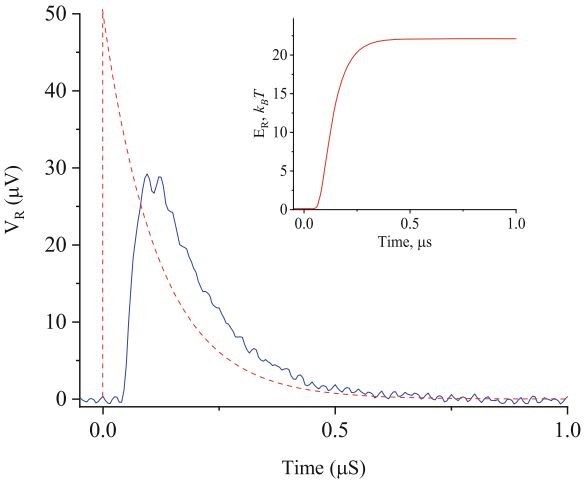

Fig. 7 The ERASE "0" WITHOUT A COPY experiment. V_R is voltage measured across the resistor. Erasure starts at $t = 0$. The voltage stored in the capacitor is 50 μV. Dashed line: $V_R = 50\mu V_0 \exp\left(-\frac{t}{0.11\ \mu s}\right)$ Inset: energy dissipated in the resistor $R = 1.1$ kΩ in an ERASE WITHOUT A COPY operation

To summarize the experiment, we would like to point out that the obtained results are in no way unexpected—they are in complete agreement with any elementary physics textbook, even though the measurement of energies much less than k_BT can be challenging. In spite of that the subject of minimal energy dissipation required for computation remains a controversial topic. Curiously, in recent publications [17–20] it has been claimed that in all possible electron charge-based devices the minimal energy value k_BT ln2 must be multiplied by N, where N is the number of charge carriers (electrons or holes) representing a bit of information, typically cited as 10^4 [19]). Based on this assertion, several far-reaching conclusions were drawn about a supposed many orders of magnitude higher "energy efficiency" (by a large factor of $\sim 10^4$) of spin- or nanomagnet-based computational schemes compared to charge-based systems employing charge carrying devices (e.g., FETs) as computing elements. For example, Fashami et al. suggest in [19]:

"Transistors switch by moving electrical charge into or out of their active regions. If this process is carried out non-adiabatically, then it dissipates an amount of energy equal to at least NkT ln $(1/p)$, where N is the number of electrons (information carriers) moved into or out of the device, T is the temperature and p is the 'bit error probability' associated with random switching [21, 22]. On the other hand, if logic bits are encoded in two stable magnetization orientations along the easy axis of an anisotropic single domain magnet (or the single-domain magnetostrictive layer of a multiferroic nanomagnet), then switching between these orientations can take place by dissipating only $\sim kT$ ln$(1/p)$ of energy, regardless of the number of spins (information carriers) in the nanomagnet [22]. This is a remarkable result and accrues from the fact that exchange interaction between spins makes all the $\sim 10^4$ spins in a single-domain nanomagnet behave collectively like a giant single spin [22, 23] (a single information carrier) and rotate in unison [22]. As a result, for the same bit error probability p, the ratio of the minimum energy that must be dissipated to

switch a nanomagnet to that dissipated to switch a nanotransistor will be $\sim 1/N << 1$. This makes the nanomagnet intrinsically more energy efficient."

This assertion is incorrect, and to trace how it came about one needs to take a closer look at its first appearance in a paper by Salahuddin et al. [22]. To clarify this topic, we must make an examination of some of the statements made in reference [22]. In the introductory part of [22], an in-series RC circuit, where a capacitor C is charged and discharged to a voltage V, is considered as a bit processing element. It is stated in [22] that: "... the discharging process involves charging the next stage, making the total dissipation in one cycle equal to $CV^2 = NqV$, where N is the number of electrons and q is the electronic charge." One can immediately note two issues. First, the expression $CV^2 = NqV$ is nothing but an identity, simply based on the notion $Q = CV$, while for the circuit considered in [22] charge is a continuous variable, and unless effects of charge quantization are in play there is no requirement on the discreteness of charge. Second, the assertion: "... the total dissipation in one cycle equal to CV^2" is only true in a special case of square wave with a period $T > RC$. Indeed, energy dissipation per cycle can be much greater than CV^2 if $T << RC$, to the point when no energy is delivered to capacitor and all of it is dissipated in the resistor, or much less than CV^2 if charging is performed quasi-adiabatically as we showed above.

Next, the key derivation for the claim $E_{\min} = N\ k_B T\ \ln 2$ is presented in [22] as follows: "It can be shown that for an error probability of $1/r = I_{\text{off}}/I_{\text{on}}$, thermodynamics requires the minimum voltage to be $V = k_B T/q\ \ln r$, which translates to a theoretical minimum dissipation of $N\ k_B T\ \ln r$ or $N\ k_B T\ \ln 2$ (see Refs [1, 21] and references therein) for an error probability $1/r = 50\%$." Let us take a closer look at these assertions. First, there is no such entity as "the minimal voltage" at any given temperature unless the relevant bandwidth is specified. The *value* of $V_{\min} = (k_B T/q)\ \ln 2$ can of course be calculated and, indeed, has dimensionality of volts but has no special physical meaning. For example, at room temperature $k_B T/q\ \ln 2 = 18$ mV. If we now look at how "a theoretical minimum dissipation of $N\ k_B T\ \ln r$ or $N\ k_B T\ \ln 2$" per bit operation is derived in [22], it becomes clear that the value of $V_{\min} = (kT/q)\ \ln 2$ was inserted in the *identity* $CV^2 = NqV$ (which, as we showed above is, strictly speaking, incorrect for the considered case of an RC chain charged from voltage source) *and then equated to what is claimed to be a minimal energy:* $E_{\min} = Nq\ V_{\min} = Nq(k_B T/q)\ \ln 2$ resulting in $E_{min} = N\ k_B T\ \ln 2$. This derivation undoubtedly is an example of circular reasoning, a logical fallacy in which the reasoners begin with what they are trying to end with [24]. *No other proof for this statement for a minimum energy could be found, neither in the text of* [22] *nor in the references cited in that paper*. It should be noted that the result obtained in [22] for the minimum energy required for non-adiabatically switching a nanomagnet, $k_B T\ \ln 2$, is correct. This, however, comes as no surprise, since the USL, the Ultimate Shannon Limit is valid for *any* state variable [2].

3 Adiabatic Circuit Approaches

3.1 Introduction to Adiabatic Reversible Systems

In order to function as a reversible computational system, the physical implementation must be both logically and physically reversible. We have seen experimentally that in agreement with Landauer's principle, it is not measurement in the system that must cause dissipation, but the destruction of information. Consequently, one can make copies of a bit without dissipation to heat, analogous to drawing money from a bank. When use of these copies is finished, the bits can be paid back, provided that there is a record of where the bits came from (the 1 or 0 supply). It is logical reversibility that provides this record keeping. The most general implementation of a reversible computing system is shown in Fig. 8. Input information is applied to the system, left side, and the computation proceeds until the outputs are available at the right. Energy can be drawn from reservoirs to form 1 and 0 bits as needed during the computation. Once the process is completed the system enters the decompute phase, where the computation is essentially run backwards so that the bit energies drawn from the reservoirs are pushed back into the reservoirs. Control circuitry regulates the compute and decompute phases, and represents overhead circuitry needed to implement a reversible system. While conceptually simple, the circuit overhead can be significant.

A number of methods have been proposed to implement reversible computing systems. One way to achieve logical reversibility is to use logically reversible gates, such as Fredkin and Toffoli gates [25, 26], Fig. 9, where the inputs can be reconstructed from the outputs. The Fredkin gate performs a controlled swap, where if the control input C is a 1, I_1 is mapped to O_2 and I_2 to O_1, while if $C = 0$ the inputs are passed directly to the outputs. The Toffoli gate is a controlled-controlled NOT, where if both C_1 and C_2 are 1, the input is inverted and passed to the output, otherwise the input is passed directly. In addition to logical reversibility which provides the architectural information needed for energy recovery in the decompute

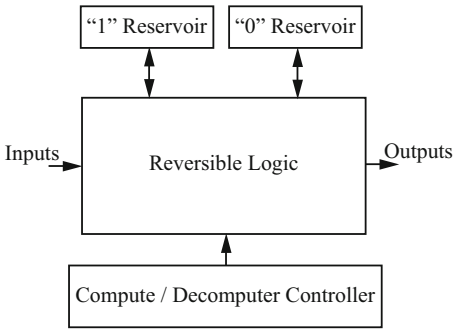

Fig. 8 Block diagram of reversible computing system

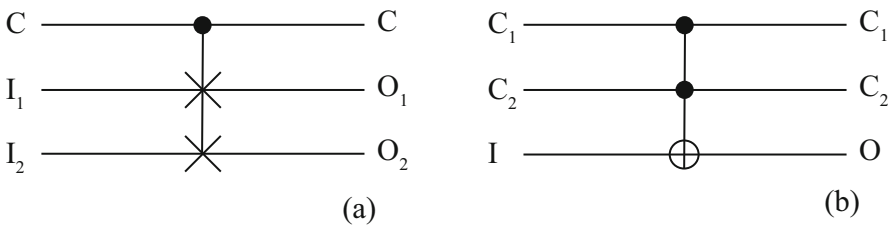

Fig. 9 Example reversible logic gates. (**a**) Fredkin gate. (**b**) Toffoli gate

phase, the system must also be physically reversible so that energy is not lost by simply operating. There must be a means for taking energy from reservoirs to form bits of information, without loss of energy, as well as a means for returning this energy to reservoirs, again with no loss of energy. Such truly reversible, adiabatic, processes are impossible to achieve in practice, so an approximation must be made with quasi-adiabatic processes.

One approach to reversible computing systems leads to circuits that can be viewed as "ballistic," where the only energy supplied to the system is that applied as the input bits. Such systems use reversible logic, and depend on the physical reversibility of the system. The bits must move "without friction" from the inputs to the outputs. For this reason, designers of circuits of this type usually point to superconducting devices as possible implementations [27, 28], and leverage progress in quantum computing experiments, an area where reversibility is a requirement. A limitation of this approach is that if energy is lost to dissipation along the way, the information may be lost, and the computation fails. If extra bits must be generated in the intermediate stages of the computation the energy for these bits must be applied as ancillary inputs, and any generated bits that are unused are "garbage" bits that must be carried along but don't contribute to the answer. The approach has the advantage of simplicity. Energy is applied only at the edges, and if the number of ancillary bits is not too large, there may be no need for a decompute stage. The only energy cost is that of the bit energies applied at the inputs.

A more flexible approach is to use low-loss reversible logic with a decompute phase. In this way energy can be taken from a reservoir to restore the energy lost to unavoidable dissipation, and to create new bits in intermediate stages of the computation. This is a quasi-adiabatic implementation that is normally referred to as adiabatic reversible logic. The degree of adiabaticity will determine the total system dissipation. As discussed in the experimental portion of this chapter, the adiabaticity is usually linked to a trade-off with speed. In an electrical system, moving energy involves moving charge through resistive components. The associated capacitance defines an RC time constant that characterizes the time scales of the system. The slower the system relative to the RC time constant, the more adiabatic the operation. The energy lost in this quasi-adiabatic operation sets the lower bound on the dissipation of the system.

Field-effect transistors have been very successful in logic applications, so it is natural to leverage the highly developed FET fabrication and circuit design expertise in the design of adiabatic reversible systems. While they might not be ideal devices, FETs provide a path toward the short-term implementation and demonstration of adiabatic systems. FETs have a number of non-idealities that make their use problematic in reversible systems, with two dominant problems. The first is the finite sub-threshold current slope [29], which leads to leakage current from source to drain. Since FETs operate by controlling the height of an energy barrier to set the current through the device, the turn-off characteristic is determined by the Boltzmann energy distribution of carriers, the so-called Boltzmann tail. At room temperature, this gives an off-current that falls exponentially below a threshold input voltage by one decade of current decrease for each 60 mV below threshold, which means that the transistor is never truly off. This finite off-current leads to dissipation that will usually set the practical lower limit on the energy dissipation in adiabatic systems. The Boltzmann tail contributes to the second major problem with FETs as devices, the finite threshold voltage. To suppress the off-current, an enhancement-mode FET is designed to have a threshold that is a few tenths of a volt, to obtain an I_{ON}/I_{OFF} ratio that is at least a few orders of magnitude. In conventional CMOS, this sets a lower limit on the supply voltage V_{DD} since the gate overdrive ($V_{DD} - V_{Th}$) determines the on-current and speed of the device. A ratio of V_{DD}/V_{Th} of 3–4 is typically used for circuits with good performance, with near-threshold or sub-threshold logic saving energy but at the cost of greatly reduced performance [30–34]. These requirements produce a web of constraints on efforts to reduce power dissipation. To limit the off-state leakage power, the threshold must be ~0.2–0.3 V. For reasonable performance, this sets a lower limit of V_{DD} of 0.7–1.0 V, which in turn makes it difficult to reduce the bit energy, $\frac{1}{2} CV_{DD}^2$ since one can only reduce capacitance through scaling. Significant research efforts in the last decade have sought devices that can break the 60 mV/decade Boltzmann limit, and cut through the web of constraints by enabling lower threshold voltages, lower V_{DD}, and consequently lower bit energies. Tunnel FETs and ferro-electric FETs have been explored as devices that can give a steeper turn-off characteristic than conventional devices.

The non-zero threshold of FETs is problematic in adiabatic circuits because it can lead to the issues of non-adiabatic dissipation and trapped energy. These problems are illustrated by the circuit of Fig. 10. Here we consider the simplest logic element, an inverter, which is logically reversible. It can be made physically quasi-adiabatic by using a single power clock with ramped edges, as shown in Fig. 10. Consider the case where the output will be set to a logic 1. A logic 0, ground, is applied to the input while the power clock is in the inactive state, ground. The desired action is for the p-channel device to be on and the output capacitor will charge as the power clock ramps up. As shown previously, the capacitor can be charged by a ramp voltage with an arbitrarily small energy lost to heat. The key is that during the charge and discharge phases the transistor must be on, and the source-drain voltage kept small. However, when the input and clock are both ground at the beginning of the ramp up, V_{GS} of the p-channel transistor is 0, and the device is off. Essentially no charging of

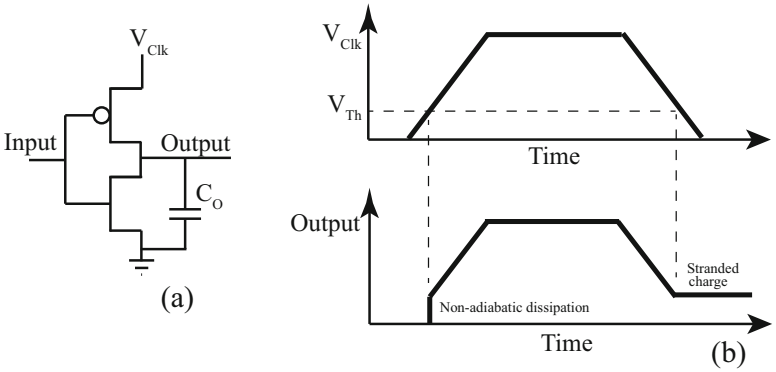

Fig. 10 (**a**) Inverter clocked with single polarity power clock. (**b**) Voltage waveforms input grounded, showing non-adiabatic dissipation and stranded charge due to FET threshold voltage

the output occurs until the voltage of the clock reaches V_{Th}, and then as the transistor turns on there is a non-adiabatic inrush of current, with $V_{\mathrm{DS}} \sim V_{\mathrm{Th}}$, resulting in energy loss. A similar problem occurs during the discharge phase. Here the output voltage starts at V_{DD}, and ramps down with the clock, with a small voltage V_{DS}. When the clock voltage falls to $\sim V_{\mathrm{Th}}$ the p-channel transistor turns off, and little charge is removed from the output capacitor during the remainder of the ramp. The associated energy left on the capacitor is trapped, and will dissipate to heat either through leakage or in the n-channel transistor if a logic 1 is applied to the input. This example illustrates the significant problems that the finite threshold voltage of transistors can cause when implementing adiabatic circuits. To overcome these issues, more elaborate logic circuits and architectures must be used, which adds to overhead associated with adiabatic logic.

Reversible logic implemented with transistors can be divided into two rough categories: Quasi-adiabatic and Asymptotically adiabatic [35]. These names can be somewhat confusing since all "adiabatic" circuits are actually quasi-adiabatic, but these labels are useful in categorizing the degree of adiabaticity. In asymptotically adiabatic circuits the dissipation has no lower limit, and is determined only by the speed of the voltage transitions. In quasi-adiabatic circuits, there is some inherent destruction of information, with unavoidable dissipation, such as the example given above. An example of quasi-adiabatic logic is positive feedback adiabatic logic (PFAL) [36], which is one of the most popular adiabatic logic families. PFAL has the advantage that it can be pipelined, and requires only a single-polarity power clock. A two-input PFAL AND gate is shown in Fig. 11a. The heart of the gate is a set of cross-coupled inverters that make up a static-RAM cell, and provide differential outputs. The pull-up network on each side of the gate determines the value that the gate will take on when the power clock is ramped up. When the clock ramp is complete, the inputs can be removed and the output state is held by the SRAM cell. Using a four-phase set of power clocks, this hold function allows pipelining of PFAL gates. A representative set of input, clock, and output signals for the gate is

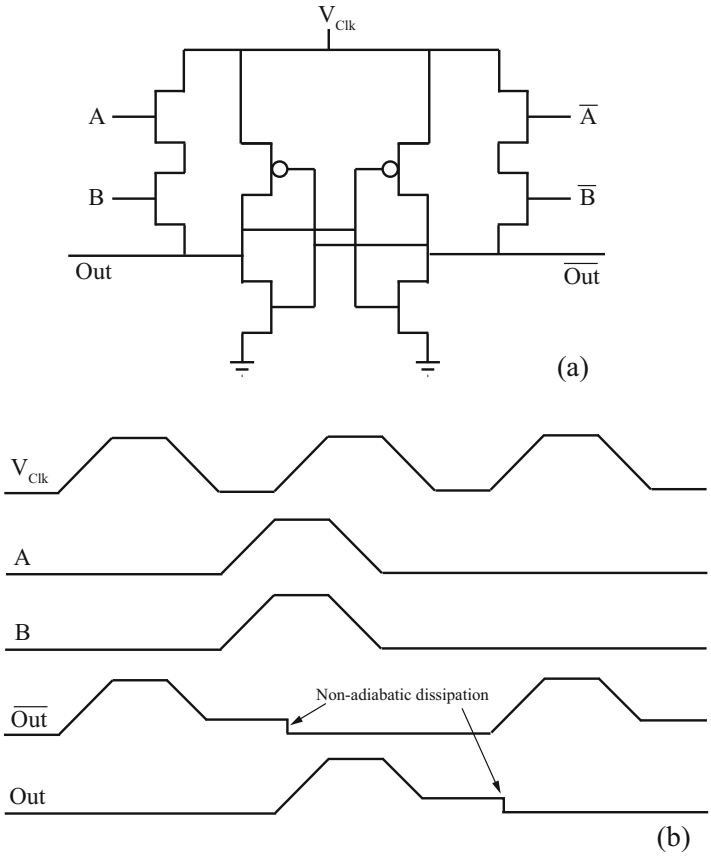

Fig. 11 (**a**) Schematic of PFAL AND gate. (**b**) Voltage waveforms showing the operation of the gate and the occurrence of non-adiabatic dissipation

shown in Fig. 11b [37]. As explained above, some energy is trapped at the output or $\overline{\text{output}}$, which causes some unavoidable dissipation.

An example of asymptotically adiabatic logic is split-rail charge recovery logic (SCRL) [38]. SCRL avoids trapping energy during clocking by using bipolar logic levels and power clocks. In an SCRL inverter source of the p-channel device is connected to a positive-going power-clock, and the source of the n-channel device is connected to a negative-going power-clock. Assigning logic 1 to $+V$ and logic 0 to $-V$ avoids the transistor threshold problem, because when an input is applied to the gate, one of the transistors (n or p channel) is on even when the clock is inactive (ground). Then, when the clock is ramped, the output capacitor begins charging immediately, so that the source-drain voltage of the transistor can be kept small, thus minimizing dissipation. Likewise, when the clock voltage is ramped down, the active transistor remains on throughout the discharge ramp, and no charge is left

stranded on the output capacitor. Pipelining is possible with SCRL, but since there is no inherent memory as in PFAL, additional circuitry is required to control the decompute, energy-recovery, phase.

3.2 *Bennett Clocking*

To recover energy from a logical system, it is necessary that any operations be logically reversible. Conceptually, the most straightforward way to achieve this is to use logically reversible gates, so that logical reversibility is guaranteed. However, it is possible to operate an irreversible logic circuit in a fashion that makes it logically reversible. An example of this is to use irreversible logic with multi-phase power clocks that form a retractile cascade, called Bennett clocking. In Bennett clocking, the power clocks sequentially energize successive levels of logic in the compute phase, and then de-energize logic in the reverse fashion during the de-compute phase, recovering the energy in the circuit. The timing of a three-level, positive going Bennett clock is shown in Fig. 12a.

Combined with adiabatic CMOS logic such as SCRL, with both positive and negative going power clocks, Bennett clocking enables a simple conversion of conventional combinational logic to reversible logic. Destruction of information is

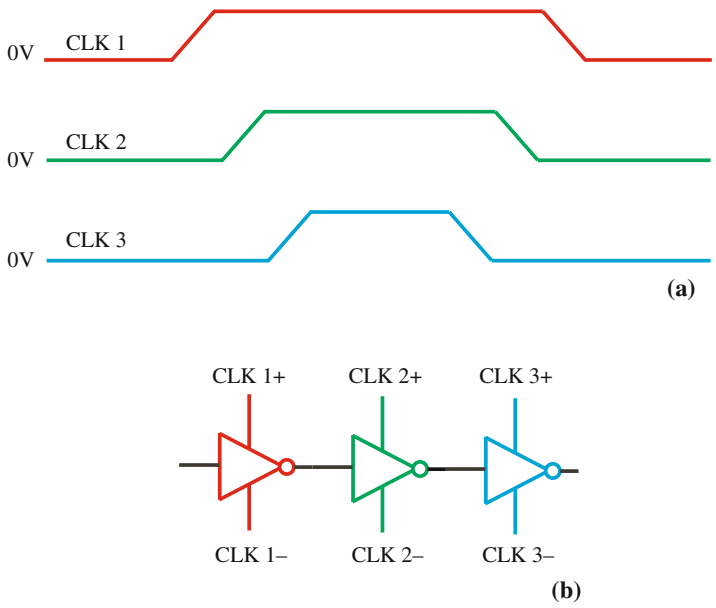

Fig. 12 (**a**). Timing diagram of three-level Bennett clocking. (**b**). Schematic of Bennett clocked SCRL shift register

avoided by retaining the inputs during the de-compute phase. In operation, an input is applied to the first gate, and then the first level Bennett clock is ramped up and held. The output from gates clocked by this level can then be used as the inputs to gates in the next. Then the second level Bennett clock is ramped, and so on. As an illustration, a three-stage Bennett clocked shift register is shown in Fig. 12b. In the de-compute, energy recovery phase, the last Bennett level, e.g. CLK 3 in Fig. 12, is ramped down, and since the inputs from previous levels are still applied to the gates in this level, the energy stored in the gates is returned to the clock as it ramps down. This is repeated until all the Bennett levels are ramped down, completing a Bennett clock cycle. Inputs to the logic block can be changed and Bennett clock ramping begun for the next computation. As mentioned earlier, energy recovery requires some overhead. In Bennett clocking this overhead is temporal, since the inputs to the logic block must be held throughout the cycle, which limits the opportunity for pipelining.

3.3 Adiabatic Microprocessor

3.3.1 Adiabatic Architecture

As a demonstration of the design of adiabatic circuits, a general-purpose microprocessor based on the Bennett clocking scheme was implemented by the group at the University of Notre Dame. This is a relatively complex circuit, which exposes many of the design and architectural challenges faced in the design of adiabatic systems, especially Bennett clocked retractile circuits. To limit the size of the project, a textbook example of a RISC mini-MIPS processor was chosen. This implements a subset of the MIPS instruction set, using an 8-bit data word length and a multicycle microarchitecture [39]. While limited in scope, this is a real-world test for a design using the Bennett clocking scheme.

Bennett clocking can recover energy in combinational logic, but a microprocessor contains a significant amount of logic that is sequential, such as latches and register files, where energy recovery is difficult. For our simple design, we did not try to recover energy in these circuits. About 60% of the total number of transistors in our implementation are in sequential elements, so the bit energies of only about 40% of the transistors can be recovered by adiabatic circuitry. However, sequential elements generally have much lower activity rates than the combinatorial blocks, so a better rate of energy recovery is expected than that suggested by the transistor count. Our simplified mini-MIPS architecture uses instruction word length of 32 bits as in a full MIPS, but our adiabatic processor implements a subset of the instructions, and the word length used in the datapath is only 8 bits. Ten instructions are implemented: addition, subtraction, bitwise AND, bitwise OR, set less than, add immediate, branch if equal, jump, load byte and store byte; but these are enough for universal computation. Details of these instructions can be found elsewhere [39]. The register file consists of eight 8-bit registers.

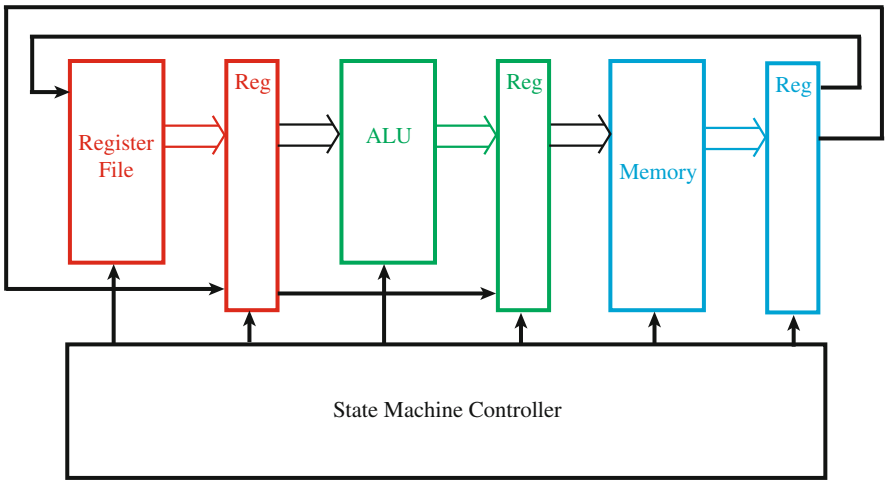

Fig. 13 Simplified top-level diagram of the microprocessor

An important design decision is the depth, or number of phases, of the Bennett power clocks. There are several important trade-offs to be considered. Each Bennett phase requires a clock generator, and the routing of clock lines to each of the associated gates. Implementing a large number of Bennett phases will therefore add considerably to the complexity of the system. However, using a small number of phases limits the design because the number of phases determines the logic depth of the combinational logic. In addition, the energy recovery efficiency increases with the number of Bennett phases because information will be destroyed in the registers separating Bennett blocks. More Bennett phases means fewer registers, and less dissipation due to data destruction. For our design, we chose to implement a 12-phase Bennett clock system.

A top-level block diagram of our RISC implementation is shown in Fig. 13. The processor interfaces to external memory, and has an internal datapath including blocks for the register file stage, and ALU. Each block is separated from the next by a register. Each logic block maps onto adiabatically clocked combinational logic with one Bennett clock set. The operation of the processor is controlled by a state machine that contains 13 states, with an additional initialize (init) state.

In our design, a block of combination logic is broken down into cascaded levels where each level is assigned to a Bennett clock phase. The Bennett clocking must then be integrated with the over-all clocked timing of the processor. In the datapath, the timing constraints are satisfied by standard clocking of the registers separating these Bennett blocks. During each cycle of the standard clock, the Bennett clock runs through a full cycle of energization and de-energization of the energy recovery logic. Data is clocked into the registers when all Bennett levels are fully energized (all logic signals are valid at this point). The outputs of the registers are held at the previous values until all Bennett levels are de-energized. At this point the register

outputs change to apply the new inputs to the combinational logic block, and the next Bennett cycle starts.

There are a number of concerns that must be addressed in the design, most stemming from timing. For instance, the controller provides the ALU multiplexer select and register enable signals, which in a standard static implementation would produce valid signals practically instantaneously, limited only by gate delay times. However, in our case the Bennett-clocked adiabatic controller module produces these signals at a time determined by the logic, as well as the power-clock phase. This timing must be taken into account in the rest of the design. A relaxed or ramping control signal cannot be used in an energized datapath block, since it would produce erratic behavior that would dissipate energy. This illustrates an essential timing design criterion: control signals must be produced and applied before the controlled block can be energized by its power clocks.

This requirement leads to some important design constraints. In our processor, the same set of 12 power-clock pairs power both the control module and the datapath. The timing constraint can be met by careful design, taking into account the control dependencies of the processor architecture, as illustrated in Fig. 14. In our design, with the chosen number of Bennett levels, some control signals could not meet the timing requirements if they were computed by Bennett logic. To meet the timing, those controls had to be directly implemented as specific bits in the state register, increasing the size of the state machine. The alternatives would have been to include more Bennett levels (more power-clock pairs), or to implement the control computation partially in standard static logic, which would have dissipated more energy.

The examples of control timing illustrate the close relationship of logic design and timing in a Bennett clocked system. To ensure that the logical design and timing are both correct goes beyond the capabilities of current computer-aided design (CAD) tools. For this reason, we developed a suite of tools to aid in the design of the processor.

3.3.2 Adiabatic CMOS Design Tools

Since CAD tools do not currently exist for adiabatic CMOS, we developed computer tools to aid in the logic design, design verification, and design automation of adiabatic CMOS circuits. While these tools can be applied to some other styles of energy-recovery logic, they were designed specifically for Bennett Clocked Adiabatic CMOS circuits using SCRL. The main benefit of this approach is the high degree of compatibility with the standard CMOS design flow and fabrication process, while achieving the recovery of most of the signal energy.

The main components of the developed software tools are:

1. Ramp Logic Semi-Timing Simulation Environment.
2. Bennett Energization Sequence Checker.
3. Standard Logic Synthesis Integration.

Fig. 14 Spreadsheet showing the timing dependence of some of the processor subsystems

Ramp Logic Timing Simulation Environment We built upon industry standard logic level design tools to implement structural modeling and semi-timing simulation of the Bennett clocked adiabatic circuits. The hardware design language (HDL) chosen was *SystemVerilog,* and we used the simulation environment *Mentor Graphics ModelSim*.

The signal/circuit models are discrete-level and discrete-time, but are still able to address the important timing constraints of the retractile circuits. The tools consist of two components: Bennett Wrappers Package, and the Bennett Gate Model Library.

Bennett Wrappers Package The first issue that must be addressed in simulating adiabatic logic is the presence of additional logic states. The Bennett wrapper makes this extension and contains the ramp logic signal type definition along with signal generator and conversion functions. While the standard Verilog/SystemVerilog logic signal type has four states [1, 0, X, Z], the adiabatic circuits spend a significant time in transitional states, so they require verification using a model which includes these transitional states. The new ramp logic signal type has nine states [X, RLXD, ACT1, ACT0, REN1, RDE1, REN0, RDE0, Z], including a relaxed (or "null") state, as well as separate energization/de-energization transition states for both logic 1 and 0. In a semi-timing simulation, each signal transitions through the intermediate states when there is a switching event. This requirement allows us to check for behaviorally accurate and correct energization of the structural components.

Bennett Gate Model Library In the design of the processor, we developed a CMOS standard-cell library containing 45 logic gates, with physical layouts, and modules describing the behavior of the logic gates with Bennett power-rail modifications. While a standard minimal logic gate model defines only the mapping between the input and output signals, a model for a Bennett clocked gate must also treat the power-clocks as inputs, and define the behavior based on the energization level of all inputs. These are modeled using the newly defined 9-state ramp logic type. This ensures that a gate model will produce the valid logical output only with proper timing and the energization of the inputs and the power-clocks. The proper timing of inputs and power-clocks is of paramount importance in Bennett clocked circuits, where signals must be applied in a well-defined order. Figure 15 shows the beginning of a processor execution test program, showing the repeating activation sequence of 12 Bennett levels while executing the test program.

Bennett Energization Sequence Checker The use of a standard gate model library for this Bennett-clocked circuit ensures that the logical function of a design is fully-specified, building on the designer work, also correct. However, it does not guarantee that the circuit has been connected to the power-clocks so that the adiabatic energy-recovery operates correctly. Our design, like any general Bennett-clocked design, is a mixture of static and adiabatic logic. It is sometimes necessary to include circuit elements that have the power-clock inputs connected to static V_{DD} and V_{SS} rails, so that it operates as standard CMOS with the associated loss of all signal energy. However, the main goal of the design effort is to maximize the adiabatic recovery and use as few standard statically powered blocks as possible.

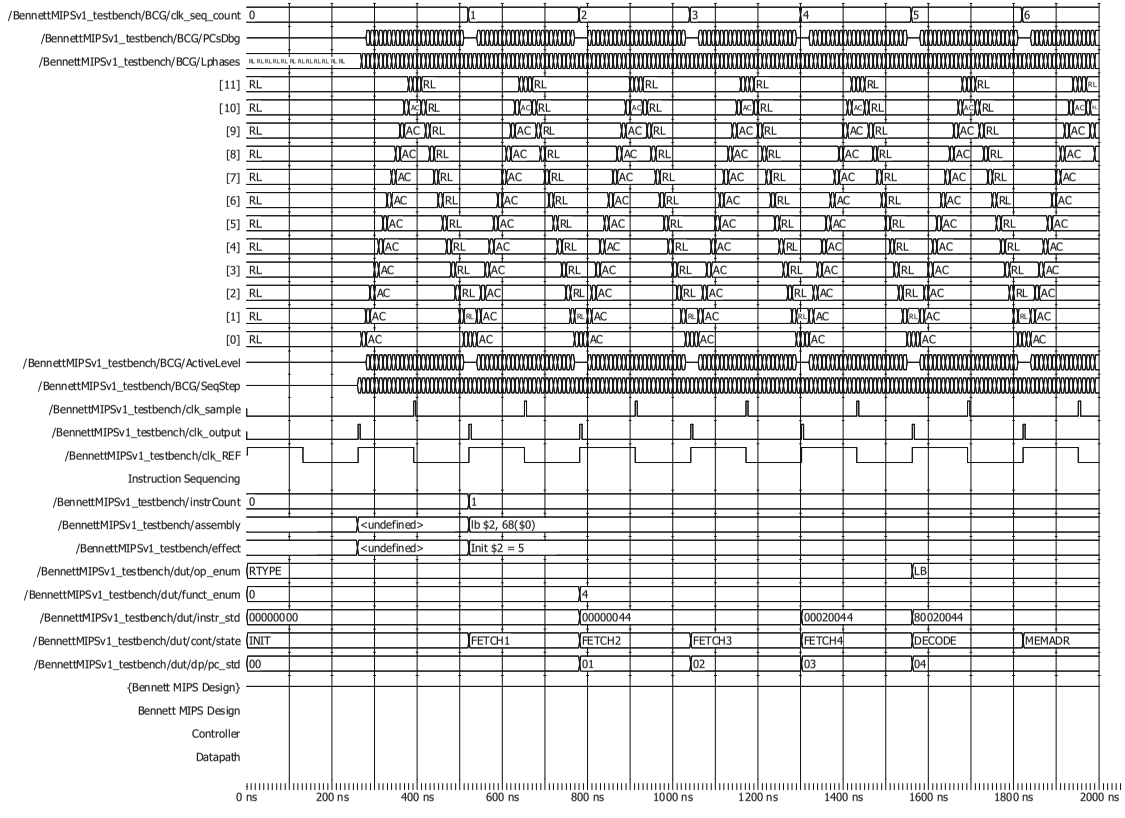

Fig. 15 Screenshot of the beginning of an execution test program

Therefore, the semi-timing simulation should perform automatic checking of the adiabatic energization sequence in every node in which the designer intends to achieve energy-recovery, and correct interface to any static logic.

The Bennett Energization Sequence Checker contains the SystemVerilog assertion functions/tasks and compiler macros to check the adiabatic charging and discharging of the specified circuits nodes as defined by the logic designer. The macros are typically called in the definition of a new adiabatic module, where they check the specified signals and generate debug output as necessary. Every new adiabatic module must contain at least two assertions: First, all incoming power-clocks, since these are used for internal consistency, and second, all incoming data signals, so that they can be checked for validity during the specified time interval.

The power-clock consistency checker has a vector input port for all positive power-clocks and all negative power-clocks. To ensure that the Bennett Clocks are valid, the checker produces an assertion error if any power-clock pair is not correctly complementary, or if any phase transitions at the wrong time. This would be a transition of any power-clock below the current highest active level, since all lower levels should be stable at the active level.

The signal sequence assertions are used to check that a given set of signals energizes in the correct order, and that there are no transitions or illegal states below the highest active level. For example, an adiabatic module requires that all data input signals must have a stable active logic value (not relaxed or ramping), before the lowest power-clock of that module begins to transition.

Standard Logic Synthesis and Integration The split-rail Bennett circuits can be synthesized using a standard CMOS tool flow with only minor modifications. By implementing adiabatic logic with gate and logic level compatibility, the approach brings energy benefits over standard CMOS circuits even in the near-term. Figure 16 illustrates the design process, which has been preliminarily tested in a standard CMOS design environment from the Cadence Design Systems. The current version of the Bennett interfacing tool (a C++ standalone netlist parser program) is in an early alpha stage, with only the extraction of structural netlist and identification of Bennett levels implemented.

Design Entry and Logic Synthesis The design entry proceeds using, for example, structural or behavioral specification in a hardware description language (HDL). A standard logic synthesis tool produces a structural gate-level netlist, based on the gate library characteristics. Since the transistors in our approach are already sized and the logic gates constructed exactly as in the standard CMOS, all automatic sizing optimizations and balancing of the delays in the combinatorial networks are directly valid.

Bennett Placement Constraints The Bennett placement constraints ensure that the standard gates will be placed into the physical part of the floorplan where they can be efficiently wired to the correct power-clocks. For each instantiated gate, this is determined by the relative logic level in the Bennett clock cascade. The

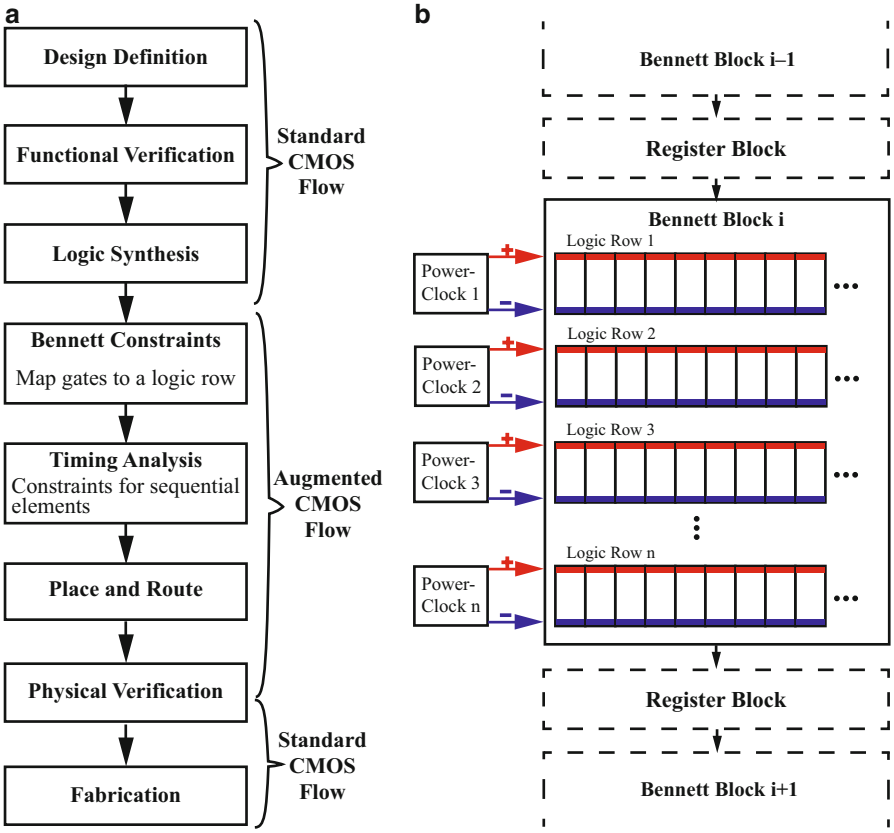

Fig. 16 (**a**) Design process for Bennett-clocked adiabatic circuits. (**b**) Mapping of logic design onto physical layout using standard cells

floorplanning can be very flexible, in principle, but in our initial design, we chose to implement the most straightforward approach to place each specific level in a logic block to one physical row. Inside one Bennett block, each pair of power-clocks is driving only one row, which simplifies the wiring complexity to approximately the same level as in the V_{DD} and ground rails of the standard CMOS. Each power-clock can also drive several separate Bennett blocks, depending on the circuit architecture choices.

The logic gate dependency information exists inside the standard synthesis tool, but this information is not generally accessible from the outside, since the internal data structures are proprietary. Therefore, we decided to implement our own software, which reads in the structural netlist produced by the logic synthesis and constructs a graphical representation of the dependencies between the gates of a design. The tool basically tags each gate with a placement constraint defining in which logic row the gate is to be placed in. Figure 17 shows an example of a

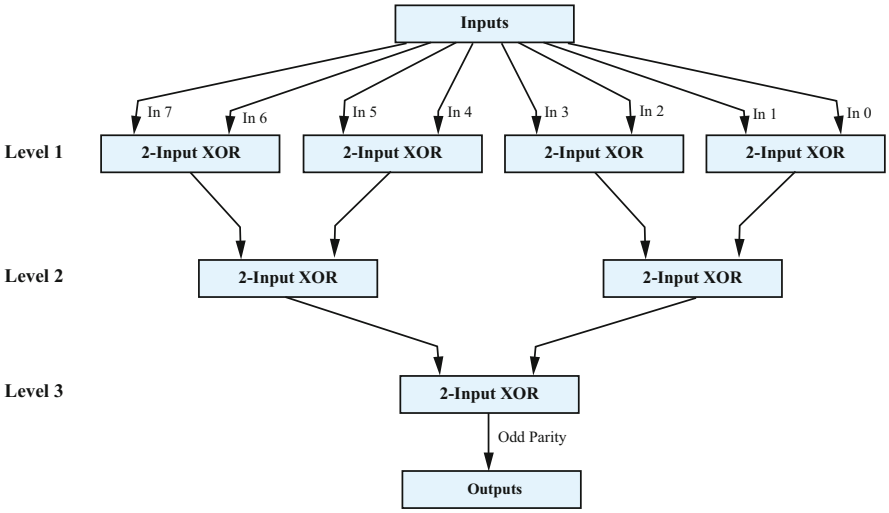

Fig. 17 Node graph for a parity generator, illustrating how a circuit is broken down into levels that are assigned to Bennett clock phases

parity generator extracted from combinatorial netlists, and placed on graph levels corresponding to the Bennett levels. This step should involve the balancing of the number of gates in each row to obtain a block with as fully utilized rows as possible, but our alpha-level tool does not yet implement this. Another important feature to be implemented is the ability to place several logic blocks together, which would improve the physical row utilization significantly.

Place-and-Route The structural netlist containing the Bennett placement constraints can be fed into a standard place-and-route tool, which constructs the physical layout of the logic part of the design and connects the standard cells with wires. The standard optimizations for combinatorial logic are valid for the adiabatic circuits. The wires for the power-clocks can then be added to drive each appropriate row of logic, for example by using the automatic functions for the clock tree synthesis. However, the power-clock routing complexity is significantly smaller than the complexity of a standard clock in a block of random logic without the clock-per-row placement constraint.

Interfacing Sequential Logic The combinatorial synthesis and placement are relatively straightforward, but accommodating the sequential elements like flip-flops and latches in the synthesized standard netlist requires considerations of the circuit timing and architecture. Basically, the standard CMOS flip-flops and latches are all compatible with the proposed approach, but their timing has to be controlled synchronously with the power-clocks. However, the location of the sequential elements in the output netlist of the standard logic synthesis has not been optimized for the retractile cascade circuits, and the best performance can be

obtained only by giving additional constraints for the standard synthesis. We have not yet implemented automation for this.

Circuit Architecture Generally, the proposed augmented design flow has relatively small overhead vs. the standard CMOS flow, but to obtain the best performance and energy offered by the Bennett clocked circuits requires some additional considerations. One of the trade-offs between the computing performance and energy dissipation is related to the size of each block and the number of power-clocks: the larger the block, the more energy can be recovered, but then a smaller number of complete computational results will be produced per clock cycle. Optimizing this trade-off is not simple because of the constraint of power density. A large block size might enable higher computation performance per clock cycle, since a small block would not be able to run as fast as theoretically possible due to power density constraints.

3.3.3 Standard Cell Design and Simulation

As mentioned in the previous section, a new standard cell library was needed for the design of the adiabatic processor. As HDL primitives typically operate only within the standard CMOS logic states, a new library was written with a behavioral description for each adiabatic gate needed in the design. These include combinatorial elements such as logic gates and transfer gate-based designs for control of data flow. Sequential elements are standard static CMOS, and thus no redefinition was needed for them. The logic design for the microprocessor was done at the gate level using this custom adiabatic library. The overall architecture is based on a standard mini MIPS, but the gate-level implementation for each module is an original design aiming to minimize logical depth to reduce system complexity and number of power clock phases.

The library includes inverters, NAND gates from 2 to 8 inputs, 2-input NOR, 2-input XOR, 2-input AND, 2-input OR, transfer gate (TG)-based multiplexers, a TG-based conditional inverter and a few complex logic modules. In addition to these adiabatic cells, cells for conventional CMOS sequential elements such as SRAMs and flip-flops were created for use in the microprocessor. A λ-based design is used to ease scaling and porting of the design to various fabrication processes.

A test-framework was created for the behavioral simulation of the MIPS processor. This framework includes descriptions of the signals needed for proper operation, such as the power clocks, and a test program [39] that exercises all functions and modules of the microprocessor. The processor was designed and verified using the adiabatic circuit design tools described previously. These simulations demonstrated that the adiabatic processor operates correctly, and during the design process, simulations of sub-systems were conducted as a debugging aid.

The transistor-level design and layout implementation of cells in the library are based on their HDL description. To verify correct operation of logic gates and modules, simulations were performed using SPICE. For each gate, a FET-level

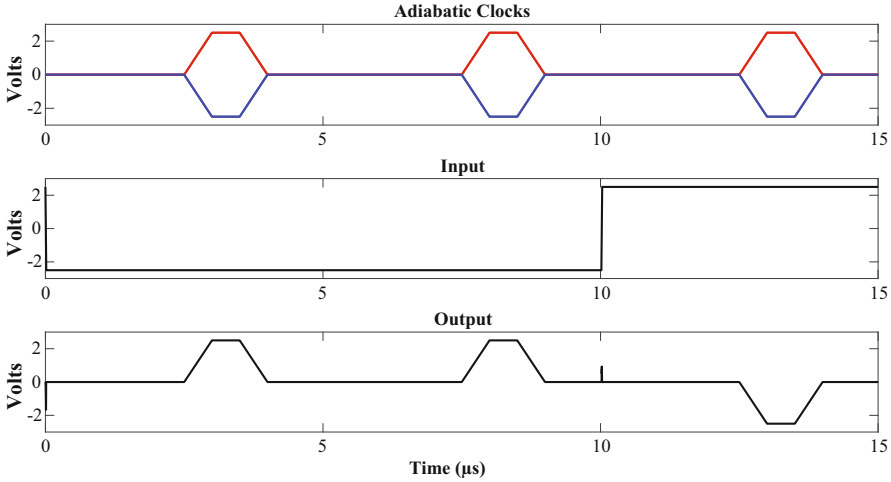

Fig. 18 SPICE simulation showing typical operation for an adiabatic inverter

schematic was created and tested by simulation before designing the physical layout. Netlist extractions, including parasitics, were performed on each of the finished layouts to ensure that the circuit design and the physical implementation match.

To simulate the adiabatic operation of the cells, tests written using SPICE directives to generate the necessary inputs and power clock waveforms. Figure 18 shows an example operation of an inverter at a frequency of 100 kHz. This shows the operation of a minimum size inverter with NMOS size 8λ and PMOS size 16λ. For this simulation, the netlist was extracted from the layout including parasitics for a MOSIS 0.5 μm process. Our simulations showed that all of the cells work as intended and follow the behavioral model.

Layouts The layout of our adiabatic MIPS microprocessor was implemented from our standard cell library. Since no synthesis tools currently support adiabatic logic, we built the entire circuit by hand, checking the layout against the HDL design. The use of a standard cell library helped in the layout of the full circuit because it enabled an orderly layout, that is easily checked against the HDL. In addition, a standard cell design is ideal for Bennett-clocked systems, since one can simply lay out the circuit so that each row contains a single logic level interfaced to a single power clock phase, which minimizes routing issues.

The standard cells were designed to comply with the design rules of a MOSIS 0.5 μm process. In addition, the design is being fabricated in Notre Dame's undergraduate IC Fabrication course using a 1 μm process. The layout uses two metal layers.

Several elements are common to all cells, and were incorporated into a template as the starting point for every adiabatic cell. The height of adiabatic cells was chosen to be 75λ, and each contains four power rails: V_{DD}, V_{SS} and the positive

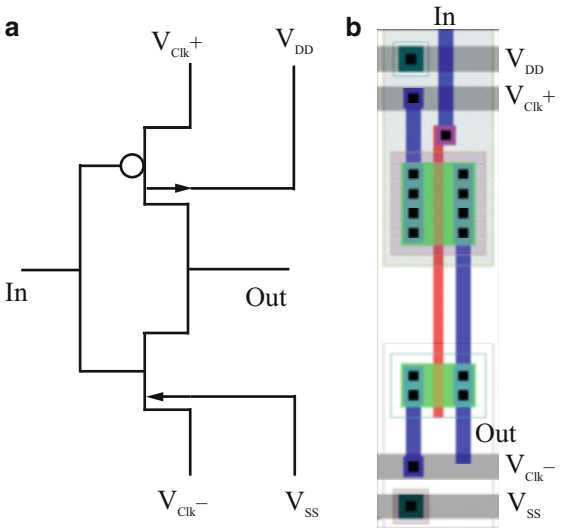

Fig. 19 Adiabatic inverter. (**a**) Electrical schematic, including substrate ties. (**b**) Physical layout

and negative power clock lines for a single Bennett clock phase. The V_{DD} and V_{SS} signals are needed to set the potentials of the CMOS substrate wells. These rails run horizontally through the cell so that when cells are stacked in a row the power rails are automatically connected. Since a row shares a Bennett clock phase, all cells in a row need to be in the same logic position in the design. By stacking rows and connecting them to sequential Bennett phases, we set the data flows from the top to the bottom of the circuit. An exception is where non-buffered multiplexers are used. Because of this directionality, we designed the cells so that inputs enter through the top of the cell, and outputs exit through the bottom of the cell. Interconnections between cells are made in the second metal layer to optimize transistor density. All layouts were created in the L-Edit software from the Tanner Tools suite. Figure 19 shows an inverter schematic (a) and layout (b).

The microprocessor layout was assembled by hand using the cells from our library. Since each row corresponds to a single power clock phase, physical placement is tied to logic design. Deviations from the rule of one row per clock phase were needed where logic levels (rows) with too many components were split into multiple physical rows to maintain a reasonable width. The data flows vertically through the design, so the rows of logic are stacked in a vertical fashion. This aims to optimize routing space for the power clocks, which contribute a large number of signals in Bennett-clocked systems. Static CMOS elements needed in the circuit have no timing requirements and can be placed in any convenient space. In our design, they were used to fill gaps and improve transistor density.

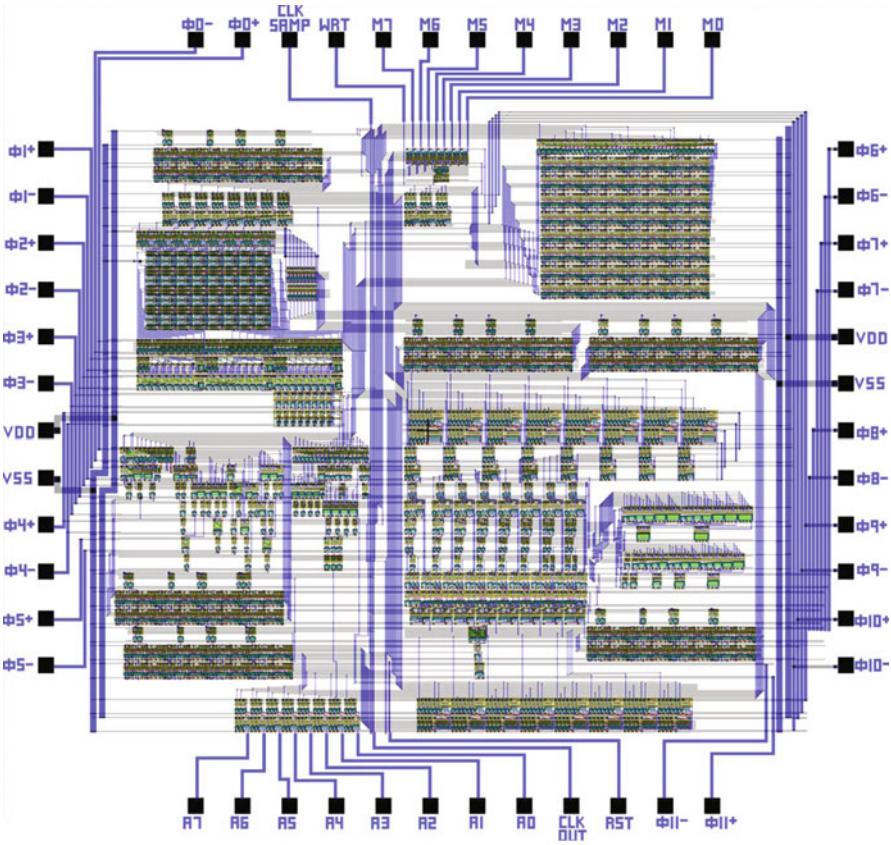

Fig. 20 Final layout of adiabatic microprocessor

The final layout for the adiabatic MIPS microprocessor is shown in Fig. 20. The adiabatic logic is located in the mid and lower portion of the layout. The upper section is dominated by the register and SRAM memory banks. Registers that are closely associated with an adiabatic combinatorial section, such as the state register for the controller state machine, are located near their corresponding logic. The total transistor count is 5766, of which approximately 40% are adiabatic.

3.3.4 Adiabatic Microprocessor Summary

Our design of the mini MIPS processor shows that significant adiabatic designs are possible. Standard CAD tools do not currently exist for adiabatic designs, so we implemented some tools to aid in the design and verification of the processor. The processor has been fabricated in the Notre Dame Nanofabrication Facility. Testing of the chip is expected to be done in the future.

3.4 New Devices for Adiabatic Logic

As discussed earlier, in conventional CMOS, logic transistors are used as resistive current switches in pull-up and pull-down networks to control the charging and discharging of capacitors. The principal shortcoming of transistor switches is their leakage current, which leads to dissipation and sets limits on the supply voltage and bit energies. Steep devices such as tunnel FETs and ferroFets can help, but don't solve the basic problem of leakage. While adiabatic circuits can be implemented by standard MOSFETs and advanced FETs, to realize the full potential of energy recovery through adiabatic circuits will require new devices. In this section, we will touch on just a couple of the possible new devices that could be well-suited to adiabatic reversible logic: adiabatic capacitive logic, and quantum-dot cellular automata.

3.4.1 Adiabatic Capacitive Logic

Nanorelays have been investigated for use in logic since they eliminate the leakage current, and in principle their operation can be scaled to close to the practical limit ($\sim$100 k_BT) for irreversible operations [40]. Unfortunately, nanorelays are not a good fit to the conventional logic paradigm due to strongly conflicting operating requirements. Since the device current flows through mechanical contacts when closed, these contacts must have a very low resistance, which suggests a large contact area and soft materials such as Au. However, to reduce the switching energy, the relay and hence the contact should be small as possible. Soft materials typically have higher adhesion, which leads to a greater hysteresis at disconnect and greater dissipation. In addition, switching the relay under bias, as in conventional logic, degrades the contacts due to arcing. This contributes to the greatest challenge facing nanorelays, endurance, since contact degradation limits the lifetime. For a device lifetime of 10 years, an endurance of at least $1x10^{14}$ on/off cycles is required but the maximum demonstrated endurance is $<10^{10}$ cycles [41]. A recent development by a group at CEA-LETI in France [42] proposes to use MEMs structures, similar to nanorelays, as variable capacitors, not switches. In this approach, called adiabatic capacitive logic (ACL), illustrated in Fig. 21, variable capacitors are used in pull-up and pull-down networks to form a voltage divider. Figure 21a shows the schematic symbols of the two types of variable capacitors used: left, one that decreases its capacitance when a control voltage is applied, and right, one where its capacitance increases with applied voltage. Implementation of these variable capacitors will be discussed below.

Figure 21b presents a schematic of a simple inverter. C_S is a variable capacitor which decreases capacitance with applied voltage, and C_L is the load capacitance, typically the input of the next gate. Additional variable capacitors can be added to the pull-up or pull-down networks to form additional logic gates. In Fig. 21c two variable capacitors are connected in parallel in the pull-up network to form an NAND gate. Since only capacitors are used in the logic circuit, there is no

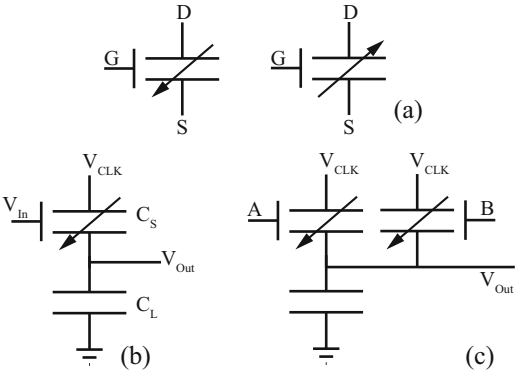

Fig. 21 Adiabatic capacitive logic. (**a**) Variable capacitor schematic symbols. (**b**) Inverter. (**c**) NAND gate

dissipation due to a direct transfer of charge from the power clock to ground. Leakage currents are essentially eliminated, and the only dissipation occurs in the parasitic series resistances when charge moves through the wires and onto the capacitor plates. Since the capacitors in ACL change value, Eq. (3) must be modified, an issue that is being addressed by the theoretical work of the CEA/LETI group, but the basic principle remains: ramp times greater than the RC time constant yield energy savings, and since no transistors are involved R can be made small by using low resistance materials.

In the operation of the inverter, when an input "1" ($V > V_{\text{inLow}}$) is applied to the variable capacitor, its capacitance decreases, so that when the adiabatic power clock ramp is applied, the output voltage stays low. Likewise, if a logic "0" ($V = 0$ V) is applied the variable capacitor stays at a high capacitance and the output voltage rises to a high voltage when the power clock is ramped. The output voltage is given by:

$$V_{\text{Out}} = \frac{C_S\,(V_{\text{in}})}{C_S\,(V_{\text{in}}) + C_L} V_{\text{CLK}} \tag{11}$$

where C_L is the load capacitance and C_S (V_{in}) is the variable capacitance. This equation sets some conditions on the desired values of C_S and C_L in the inverter. For an output 0 we want $C_S < C_L$ and for an output 1 we want $C_S > C_L$. These conditions can be met by the spacing and areas of electrodes. Preliminary calculations suggest values of $C_S = 10\ C_L$ for an output logic 1 and $C_S = C_L/4 =$ for a logic 0. These values should be possible with the devices described in the fabrication section. It should be noted that energy is supplied to the circuit from the ramping power clock, so the gates can demonstrate the power gain needed to provide noise immunity, cascadability, and drive fan-out. The fact that a capacitively coupled clock can deliver energy to a circuit and provide power gain was demonstrated experimentally in a single-electron device that shares some similarities with the proposed circuits [43]. Since C_L is increased in a fan-out structure, the driving gate

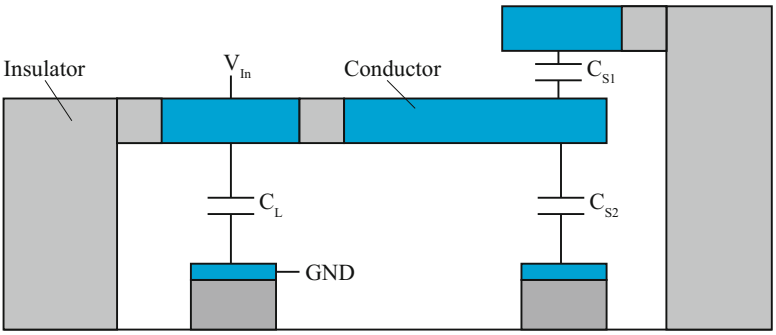

Fig. 22 Notional diagram of voltage-controlled capacitor

must be specifically designed to retain the appropriate capacitance ratios, but power gain using the clock as the power source makes fan-out possible.

A notional diagram of a MEMS, nanorelay-like structure of a voltage-controlled variable capacitor is shown in Fig. 22. Here, the position of a cantilever is controlled by an input voltage V_{in} across the capacitance C_L, while the electrically isolated end of the cantilever forms the capacitances C_{S1} and C_{S2}. With no input voltage applied, the end of the cantilever is close to the top electrode, making C_{S1} much larger than C_{S2}. When a voltage is applied to the input capacitor C_L the cantilever moves down, increasing the value of C_{S2} and reducing that of C_{S1}. Since no current is required to flow through any contact, the contacts can be coated with an insulator, and the problem of leakage current is eliminated.

Adiabatic capacitance logic operates by moving charge on the output node between the plates of C_S and C_L as the power clock is applied. Depending on the value of C_S the output voltage takes on either a high or low value. The lack of a direct electrical connection to the output node is an advantage in that it essentially eliminates leakage currents (except for low-level effects such as tunneling) and the associated dissipation. However, the fact that this node is "floating" makes it vulnerable to small leakage currents such as surface leakage, and tunneling that can, over time, charge the node and shift the output voltage. To mitigate this effect, it is possible to take advantage of the fact that the MEMS capacitor structures and logic gates can operate with both positive and negative power clocks. When the gates are operated with an alternating series of positive and negative clocks, any parasitic leakage will cancel out over a full cycle, keeping the floating node at no net charge. While the leakage might be data dependent, meaning that the leakage is different if the bit is a 1 or 0, the alternating clock voltage should keep the floating node close to 0 V. There is little impact on circuit and architecture design since the sign of neither the input nor the clock affects the operation of the gate. Where conversion to conventional logic such as CMOS is required, a simple absolute-value circuit can be used.

Figure 23 illustrates the operation of an ACL shift register, a small circuit demonstrating the principles of adiabatic operation. The shift register is driven by

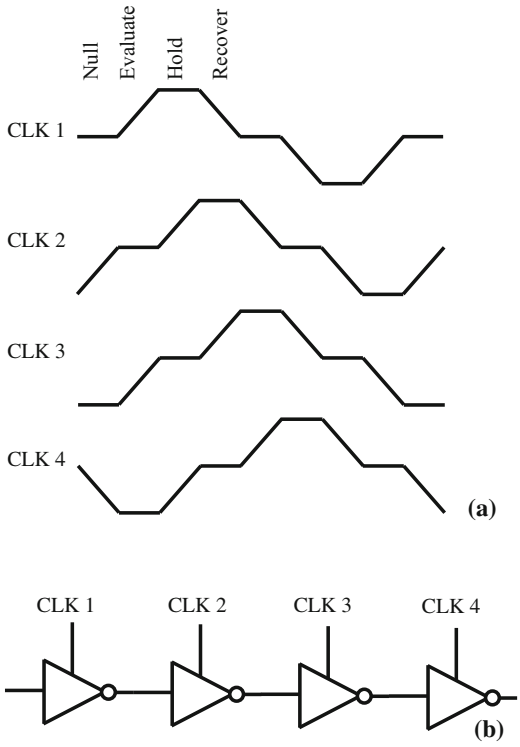

Fig. 23 (**a**) Four-phase bipolar clock. One period of each phase consists of two sequences of Null, Evaluate, Hold, and Recover. (**b**) ACL shift register

a 4-phase power clock, shown as flat-topped ramps, where each phase is shifted by 45° from the previous phase, Fig. 23a. Each half-phase of a clock (first positive, then negative) can be broken into four phases: Null, Evaluate, Hold, and Recover. Consider the first inverter shown in the shift register, Fig. 23b, powered by CLK1. The input is applied during the Null phase. Next, CLK1 ramps up (Evaluate) and the proper value appears at the output, which is applied to inverter 2. While inverter 1 is in the hold phase, CLK2 ramps up for the Evaluate phase of inverter 2. Next, CLK1 ramps down (Recover) and energy is transferred from the inverter back into the clock. New data can be applied to inverter 1 as the old bit shifts down the line. Note that either a positive or negative voltage can represent an input logic 1, regardless of whether the clock will go positive or negative.

3.4.2 Molecular Quantum-Dot Cellular Automata

The scaling of FETs becomes very difficult as device dimensions approach a few nm. This can be seen in the naming of the latest technology nodes. For instance, in

14 nm node devices the gate length is approximately 20 nm, and the node name is just that: a name. The difficulty is that at these dimensions quantum mechanical effects, especially tunneling, become significant. Since the off-state of an FET should have no current, it is necessary to suppress tunneling, while the on-state requires a high current flow. It is very difficult to meet both of these requirements at dimensions of a few nm. Size scales of a few nm bring one into the realm of molecules. Molecules as electronic devices are attractive because the structure of the molecule can provide energy barriers that are very high, with spatial modulation lengths of less than a nm, promising functionality that is not possible in top-down fabricated structures. The bottom-up fabrication of molecules is attractive because chemical synthesis takes care of the assembly of atoms on the sub-nm scale.

Over the years, there have been a number of attempts to use molecules as electronic devices [44–53], but these have mostly been attempts to make molecules into devices that resemble traditional current-carrying electronics. The problem is that small, individual molecules do not like to carry macroscopic currents, and the contacts to the molecule generally dominate the device operation. There is another device paradigm, quantum-dot cellular automata (QCA) that proposes to use molecules as charge containers rather than conductors. This builds on the extensive work on QCA, a computing paradigm that was developed and first demonstrated at the University of Notre Dame [54–64]. Operating QCA cells and small circuits have been demonstrated in a number of implementations, including metal dots [59], semiconductor dots [65], magnetic domains [66], and dangling surface bonds in Si [67]. What makes molecular QCA particularly interesting is that molecules represent the ultimate size scaling, and QCA maps well onto adiabatic reversible computing.

The basic premise of QCA is to encode information in the position of charge (electrons or holes) within the mixed-valence molecule, which is the basic element, called a cell. The charge can move within the molecule via tunneling, but cannot leave the molecule. Figure 24 shows a notional diagram of a molecule where there are three localization centers, or dots, within the molecule. In the actual molecule, these localization centers can be a metal atom surrounded by ligand and other atoms, but the actual atomic structure of the molecule is omitted here. The synthesis of such a molecule is the subject of on-going research [68–75], and some initial demonstrations of controlled charge switching within a molecule have been made [76, 77], but the topic is beyond the scope of this chapter. The molecular cell is in

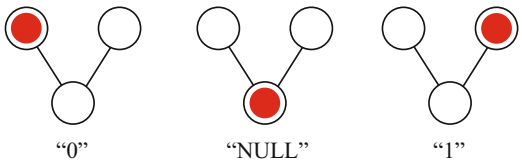

Fig. 24 Notional diagram of three-dot molecular QCA cells, showing the three polarizations 0, NULL, and 1. The polarization of the cell is determined by the location of the mobile charge

the logic 1 state when the mobile charge, say an electron, occupies the left dot, the logic 0 state when the right dot is occupied, and a NULL state when the bottom dot is occupied. If there is no input the occupation probability of each of the dots should be equal, so that the three states are energetically degenerate. If an input electrode applies a potential to the dots of the cell, the degeneracy of the states is broken, and the cell is polarized into one of the states. In particular, the NULL state is important because it will enable adiabatic clocking of the cells, where a power clock is applied as an input to the NULL dot. When energized, the electron is pulled into the NULL dot. An input potential can then be applied to select either the 0 or 1 dot, and as the power clock is relaxed and the cell enters the active state, the electron will tunnel to the dot selected by the input. Clocking the cell enables a molecular cell to exhibit power gain, where a weak input applied to a cell is brought back to full strength by the cell and then passed to the next cell. Power gain has been demonstrated in metal-dot QCA [43].

In a QCA system, cells are physically placed to implement the desired computational function. The cells are coupled by Coulomb interactions, but do not exchange charge, so that the state of one cell can act as the input of a neighboring cell. Figure 25a shows a molecular system suitable for computation. The molecules are attached to lines on the surface of the substrate, above buried clocking lines that control the flow of information. Note that the clocking lines do not need to address each individual cell. As a simplified example of how a QCA system works, Fig. 25b shows a "top view" of an array of cells in the active state, illustrating how an array of QCA cells can be arranged as a line to transmit information, and as a majority gate, Fig. 25c. A majority gate takes three input polarizations and produces an output polarization that is the majority vote of the inputs. A majority gate can be used as an AND gate if one of the inputs is set to a 0, or an OR gate if an input is set to a 1. Cells can also be arranged to make an inverter. Majority gates and inverters represent a universal set of combinational logic, and clocked QCA cells can also

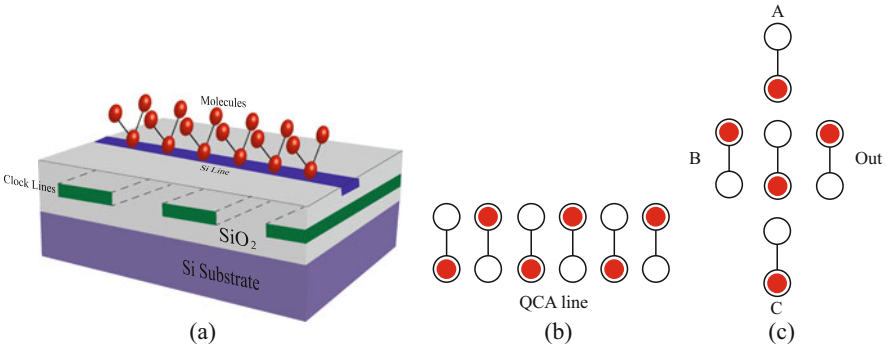

Fig. 25 (**a**) Molecular chip showing buried clock lines and silicon molecular attachment lines. (**b**) Active cells in a QCA line. (**c**) QCA majority gate

implement sequential functions such as latches. Thus, any computational system can be implemented using QCA.

The greatest difficulty facing molecular QCA systems is that the interfacing of molecules to top-down fabricated structures is difficult. The mixed valence molecules that have been produced so far measure 2–3 nm across, making the alignment of input and output electrodes difficult. Our preliminary calculations show that single-electron transistors are capable of measuring the position of an electron within a molecule, but an experimental demonstration remains to be done.

4 Direct On-Chip Measurement of Dissipation in Adiabatic Systems

The measurement of power dissipation on a chip for irreversible dissipative logic is typically obtained by conventional electrical techniques (i.e., by measuring averaged currents and voltages supplied to the chip). However, power measurement of adiabatic switching circuits becomes a complicated task, not only because the adiabatic circuits are powered by time-varying voltages, but also because power dissipation is masked by a large reactive power component, due to the nature of reversible adiabatic logic where charge is recycled. The first direct on chip heat measurement of adiabatic circuits was reported by Solomon and Frank [5], where a commercial bismuth telluride Peltier cooler was used as a sensitive thermal sensor. The chip was pressed onto a heat sink with a thermoelectric unit sandwiched between the chip and the heat sink and inserted into a thermally insulating enclosure to minimize thermal drift. Power to the chip was switched with a period of several seconds and a lock-in method was used for detection. The reported sensitivity in these experiments was about 3 mK/mW, i.e. 1 mW of power dissipated on chip resulted in measured temperature increase of 3 mK and minimum detectable power was on the order of 10 nW/Hz$^{1/2}$. The adiabatic circuit was tested with different load capacitors in a frequency range up to 20 MHz and yielded a square dependence on frequency with power dissipation up to ∼10 mW. The detailed heat balance of the tested adiabatic circuit was not disclosed in [5], but it is likely that the heat generated by the circuit in such a setup is not restricted to a flow through the thermoelectric converter. It also could occur through the electrical wires reaching the chip carrier and through ambient air convection. This is particularly likely if the thermal mass of the active circuit in which heat generation occurs is much less than the thermal mass of the entire substrate/chip carrier, so it acts as a tiny mass attached to a large cooled heat sink.

Therefore, we took a different approach and fabricated thin film thermocouples (TFTCs) directly atop the active elements of the Si CMOS circuits. In this case the heat generated in the circuit heats up the hot junction to a much higher temperature than in [5]. The cold junctions located away from the hot junctions in the area of Si wafer that remains in thermal equilibrium with the ambient. Due to the large

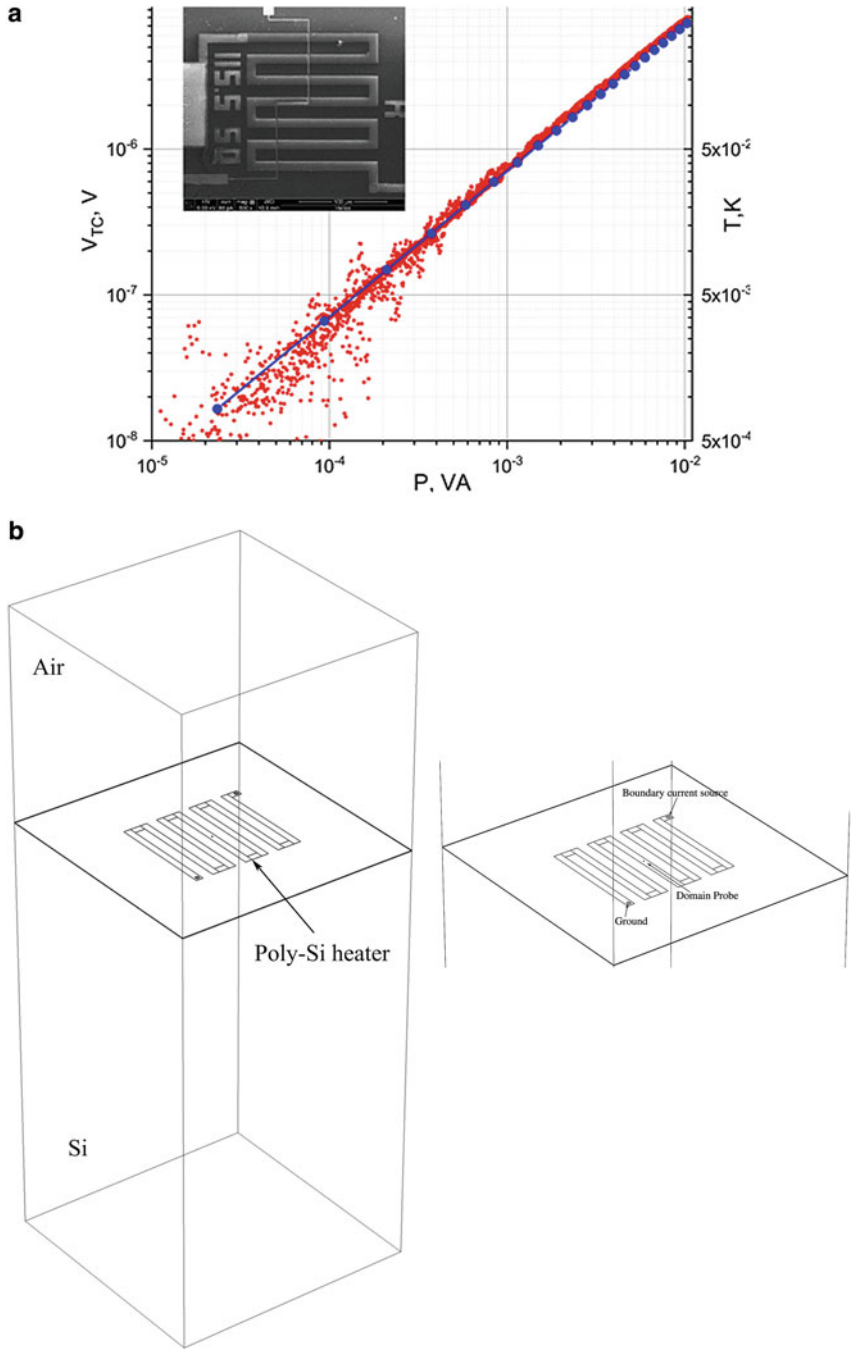

Fig. 26 (**a**) Experiment with Ni-Au thermocouple placed on top of heater separated from it by 20 nm of Al_2O_3. Experimental results (red dots) along with the COMSOL simulations (blue dots)

thermal conductivity of Si, and the negligibly small thermal mass of the active circuit compared to the thermal mass of the entire Si chip, the heat sinking capability of the substrate is very large. As a result, for the range of power dissipation discussed below (<10 mW) the temperature at the surface of the chip quickly drops to the ambient temperature (∼293 K) for distances >10 μm away from the powered circuit.

We fabricated 1 μm wide Ni-Au TFTC with the hot junctions directly atop the active elements, separated by a thin (∼200 nm) layer of dielectric to provide electrical insulation. The cold junctions of the thermocouples are located about 50–500 μm away to ensure minimal heating. To calibrate this test-bed and determine the limits of measurable power using this approach we used TFTCs fabricated on top of ∼1.15 mm long, 10 μm wide, and 200 nm thick poly-Si resistors, with hot junctions having dimensions $1 \times 2\ \mu m^2$ situated in the middle of the heater, inset in Fig. 26. Experimental results are compared with the simulations from COMSOL Multiphysics software using the electric current and heat transfer modules. The temperature increase at the center of the wire was calculated using COMSOL's domain probe, and the simulated temperature is shown in Fig. 26a along with the experimental results. The simulations take into account the heat flow along the heater and into the air and substrate. The structure for the simulations, Fig. 26b is matched to the physical layout of the experiment, inset Fig. 26a, i.e. the 200-nm-thick poly-Si heater is placed on top of a 600-μm-thick Si wafer with 400 nm thermally grown SiO_2. The heater is coated with 300 nm SiO_2 and 25 nm ALD deposited Al_2O_3. In order to simulate the Joule heating of the poly-Si heater, one end of the wire was grounded and current was applied through a boundary current node. The applied current was varied between 0.05 mA and 1.05 mA to match to the applied power in the experiment. The experimental results shown in Fig. 26a are obtained using two techniques[7]: one, the so-called "2 ω" method when a sinewave at a frequency ω is applied to a heater which results in a heat signal oscillating at a frequency 2 ω producing an electrical response in a thermocouple at the same frequency 2 ω and then measured by a lock-in amplifier (SR830). The resulting signal is then multiplied by $\sqrt{2}$ to account for the built-in RMS coefficient in the lock-in amplifier measurement [78]. Alternatively, a triangle wave from 0 to 10 V is applied to poly Si wire leading to the oscillation in power dissipation in the entire heater structure from 0 to 10.5 mW. The TC response is acquired with a digital oscilloscope (Picoscope 6404) and averaged to improve SNR. For the results shown

Fig. 26 (continued) showing TC response in volts (left axis) and temperature increase in K (right axis). Inset shows a micrograph of a TC on top of the meandering poly Si resistor. The Seebeck coefficient of $S = 20\ \mu V/K$ was used for conversion. (**b**) Structure and close-up of the model of the poly-Si heater attached to a Si substrate with a domain probe representing a thermocouple hot junction atop the structure

[7]In both cases the signal from the TFTC is first amplified with a differential transimpedance amplifier (DTA).

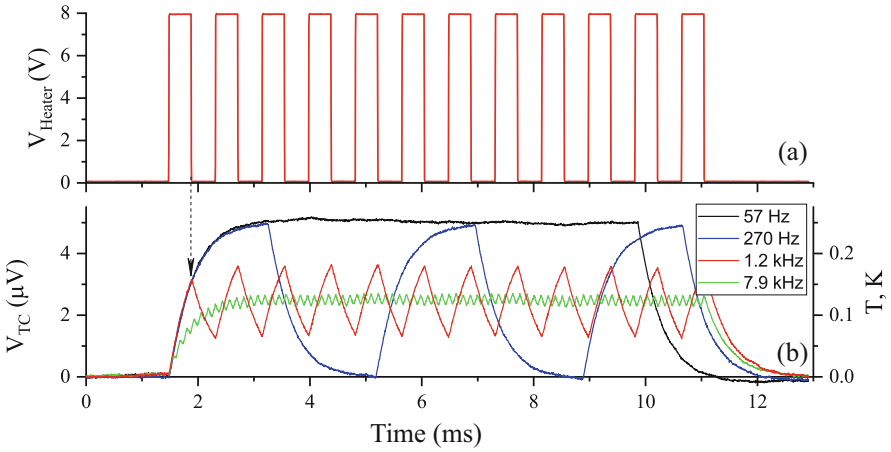

Fig. 27 Burst pulse experiment. (**a**) A unipolar pulsed signal (0 → 8 V) with 50% duty cycle is applied to the heater during the time interval from 2.5 ms to 14 ms. A 1.2 kHz train of pulses is shown. (**b**) TC response at several frequencies. At frequencies below amplifier low-pass filter cutoff (390 Hz) the temperature measured by a thermocouple reaches maximum for each applied pulse while at higher frequencies the measured TC signal corresponds to the average heating, set exactly at 50% of maximum as expected for 50% duty cycle

in Fig. 26a trace averaging resulted in $\sqrt{N} = 40$ improvement of SNR (see Sect. 2.3 for details on averaging procedure). The large spread of data at power dissipation below 0.1 mW indicates insufficient averaging. Nonetheless, both methods yielded the same results and are in good agreement with the simulations. The results show that application of 1 mW of power to the entire heater structure increases its temperature by ~40 mK. Taking into account the geometry of the heater (total length 1150 μm) and the hot junction of TFTC (about 1 μm long), the power dissipated in the segment of the heater in contact with the thermocouple is about three orders of magnitude smaller, on the order of 1 μW. From there we estimate the effective thermal resistance of this arrangement, ~40 K/mW. This is about four orders of magnitude larger than the thermal resistance measured in [5], meaning a much larger thermal signal is available for measurement. However, this advantage is mitigated by the vast difference in thermoelectric efficiency between the Ni-Au thermocouple used in this work (relative Seebeck coefficient, $S = 20$ μV/K) and Bi-Te thermopile ($S = 24$ mV/K) used in [5].

The waveforms in a digital circuit are very different from low frequency sinewaves and linear ramps, therefore another test was performed using the same test-bed. For this purpose, we devised an experiment where power and waveforms to a circuit are applied periodically at low "blanking" frequency (1–10 Hz). Figure 27 shows the results of such an experiment where the thermal signal was acquired for unipolar square-wave bursts of "filler" pulses of different length separated by blanks during which no signal is applied to the heater, Fig. 27a. At filler frequencies lower than the cutoff frequency of the ($f_T \sim 390$ Hz), the signal generated by the

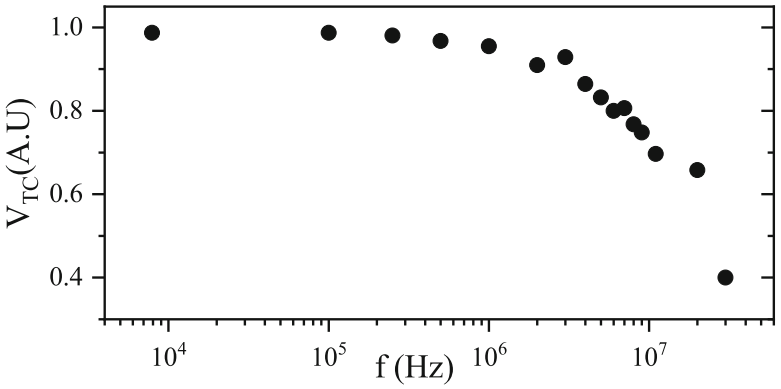

Fig. 28 Frequency dependence of thermocouple response for burst experiment. The −3 dB attenuation at 10 MHz caused by the distributed capacitance is clearly visible. Results are shown for frequencies much higher than lop-pass filter cutoff of 390 Hz

TFTC follows the shape of individual pulses so that the temperature changes from minimum to a maximum (black and blue traces in Fig. 27b), whereas at frequencies above f_T, the *average* temperature is measured (green trace in Fig. 27b), resulting in exactly half of maximum temperature for a 50% duty cycle. The experiment was repeated for filler burst pulses up to 20 MHz. The resulting frequency dependence of TC signal is presented in Fig. 28, where power dissipated in the resistor remains constant from several kHz to approximately 2 MHz. Above that frequency the distributed capacitance of the poly Si resistor to the substrate through a 400 nm thick layer of SiO_2 insulating the poly Si resistor from the substrate (approx. 1 fF per micron length) forms a low pass filter with a cut-off frequency about 10 MHz, resulting in attenuation of power delivered to the heating element, in good agreement with experimental observations.

To build upon this work, Au-Ni thermocouples are being fabricated on an adiabatic CMOS circuit consisting of three conventional CMOS inverters connected by transmission gates for adiabatic reversible operation. Operated in reversible mode, three split rail clocking signals will be used to move charge onto the gates adiabatically, along with five signals to control the transmission gates. In irreversible mode, the transmission gates between inverters are opened (conducting) so that the three inverters are connected in series, and a single rail power supply is used. Thermocouples are fabricated directly above one of the gates of the inverters, Fig. 29, where the dissipated power density will be greatest and are electrically isolated by 20 nm of Al_2O_3 and 20 nm of SiO_2 deposited by ALD. To minimize interference from clocking lines crossed by the thermocouple leads, the length and overlap area of each lead with the clocking lines are carefully designed in the layout to be equal. Since the electrical conductivity of Au is approximately 3× that of Ni, the Ni lead of the thermocouple is 150 nm thick while the Au is 50 nm thick so that the resistance of the two leads is approximately equal. All this is to make the

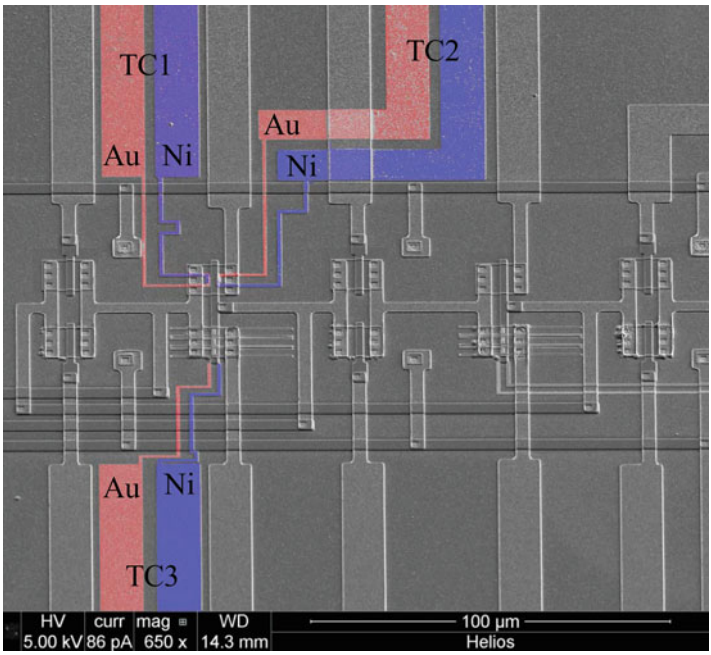

Fig. 29 A part of CMOS circuit equipped with four Au-Pt thermocouples situated above the channels of transistors

thermocouple approximately a balanced line so that any electrical noise picked up is common mode noise to both leads and can be cancelled out by the DTA used to measure the thermal signal.

Unlike in the polysilicon heater experiment, in a digital circuit most of the heat is dissipated as a transient during switching events (ignoring passive power from leakage). A finite difference simulation in MATLAB was used to estimate the thermal response and transients of a typical digital circuit. A block of silicon of size $50 \times 50 \times 20\ \mu m^3$ was simulated with a 2 μm long by 12 μm wide gate. The bit energy was estimated by calculating the capacitance of the output of the inverter using the lithographic dimensions and dielectric thicknesses, and by assuming that the dissipation occurs in the MOS channel within 1 μm of the silicon surface. Heat conduction through the poly gate and metal contacts was not simulated but is most likely a small perturbation compared to the cooling power of the bulk silicon substrate. It was also assumed that the time over which the bit energy is dissipated is comparable to the time steps of the simulation (1 ns) so that the transient heat dissipation was modeled as a pulse of energy for a one-time step re-occurring at the switching frequency. For boundary conditions, the sides and bottom of the simulated silicon are constant temperature surfaces and the top is a zero heat-flux surface. The thermal response on the channel at the silicon surface as a function of time is shown in Fig. 30. The time-averaged change in temperature is approximately

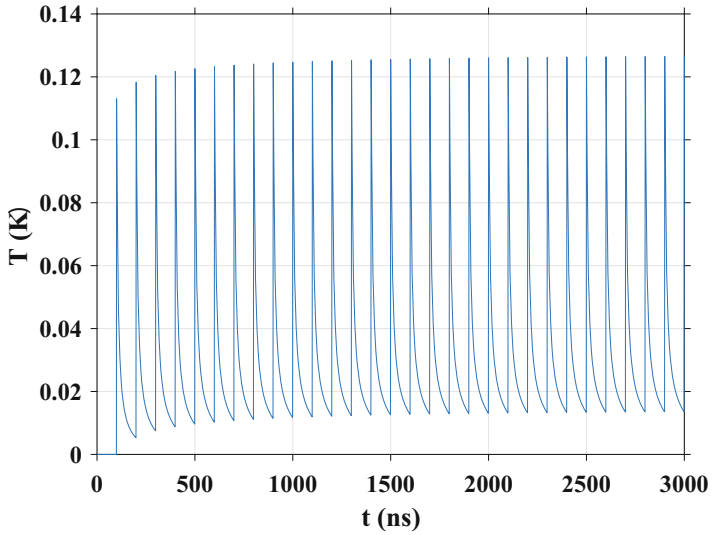

Fig. 30 Simulated thermal response of CMOS switching at $f = 10$ MHz

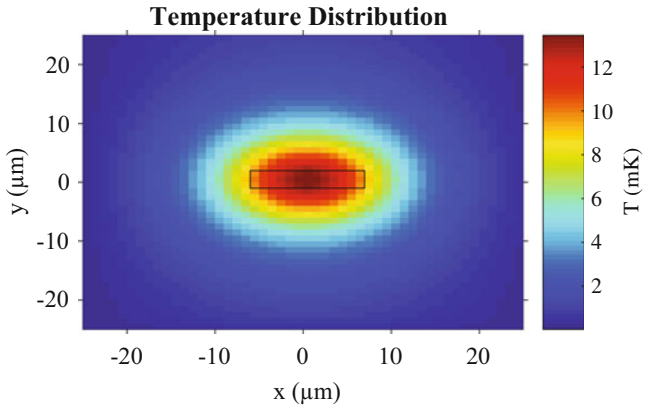

Fig. 31 Temperature distribution of switching at $f = 10$ MHz. The black box in the center shows the channel area where the heat is generated

30 mK at a switching frequency of $f = 10$ MHz and this value scales linearly with the frequency. The time averaged power is only 44 μW but this is dissipated in a much smaller volume than in the case of the polysilicon heater. Notice, however, that the peak height of the transients is determined solely by the local dissipated energy density (which is assumed to be uniform) and the heat capacity of the silicon. Because of this, the difference between the maximum and minimum temperature is constant throughout the simulation and independent of the switching frequency. Figure 31 shows the temperature distribution around the channel area at one of the

temperature minima from Fig. 30. These simulations suggest that the temperature change is well within the measurable range for on-chip thermocouples provided that the thermocouples are placed very close to the channel, and that they will provide a means of making a quantitative comparison between conventional CMOS operation and adiabatic reversible operation in the future.

5 Summary

Our experiments have demonstrated that the Landauer principle is correct and that dissipation in computational systems can be made arbitrarily small if the destruction of information is avoided. All state variables are subject to the same dissipation limit for irreversible operations, the ultimate Shannon limit. The Landauer principle applies to Boolean as well as non-Boolean computational paradigms. The only thing that matters is how information is treated in the computational system.

A significant adiabatic reversible system, a mini-MIPS microprocessor, was designed and synthesized in SCRL CMOS. This showed that new CAD tools are needed to support the design of adiabatic systems. While FETs are useful for demonstrations of adiabatic systems, the leakage and finite threshold inherent to FETs are significant limiters to their usefulness in adiabatic applications. A brief summary of two possible replacement devices, adiabatic capacitive logic and molecular quantum-dot cellular automata was given. There are challenges to both of these, but they may provide a path forward to implement large-scale adiabatic systems.

We have begun experimental measurements of CMOS adiabatic systems using thermocouples placed close above the devices. While these experiments are still in the early phases, they will provide a direct quantitative comparison of the power dissipation of irreversible and adiabatic-reversible operation.

Adiabatic reversible systems will always require a trade-off between energy savings and performance. In the past, this was deemed an unacceptable trade-off. However, with the capabilities of integrated systems increasingly limited by power dissipation, adiabatic reversible approaches offer a way forward. In fact, they may be only way around the problem of dissipation.

Acknowledgements This work was supported in part by the DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a, and the National Science Foundation under Grants ECCS09-01659, ECCS09-01659, DGE-1313583, and ECCS-1509087. The authors are also grateful to Amy L. Snider for assistance in preparation of this chapter.

References

1. R.K. Cavin, V.V. Zhirnov, J.A. Hutchby, G.I. Bourianoff, Energy barriers, demons, and minimum energy operation of electronic devices. Fluct. Noise Lett. **5**, C29–C38 (2005)
2. D.J. Costello, G.D. Forney, Channel coding: the road to channel capacity. Proc. IEEE **95**, 1150–1177 (2007)
3. W. Porod, R.O. Grondin, D.K. Ferry, G. Porod, Dissipation in computation. Phys. Rev. Lett. **52**, 232–235 (1984)
4. J.D. Norton, Waiting for Landauer. Stud. Hist. Philos. Mod. Phys. **42**, 184–198 (2011)
5. P.M. Solomon, D.J. Frank, Power Measurements of Adiabatic Circuits by Thermoelectric Technique, in *International Symposium on Low Power Design*, (ACM, San Jose, 1995), pp. 18–19
6. A. Berut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, E. Lutz, Experimental verification of Landauer's principle linking information and thermodynamics. Nature **483**, 187–189 (2012)
7. A.O. Orlov, C.S. Lent, C.C. Thorpe, G.P. Boechler, G.L. Snider, Experimental test of Landauer's principle at the sub-kBT level. Jpn. J. Appl. Phys. **51**, 06FE10 (2012)
8. H. Lu, A. Seabaugh, Tunnel field-effect transistors: state-of-the-art. IEEE J. Electron Devices **2**, 44–49 (2014)
9. D. Patterson, The trouble with multicore. IEEE Spectr. **47**, 28–32 (2010)
10. H. Esmaeilzadeh, E. Blem, R. Amant, K. Sankaralingam, D. Burger, Dark Silicon and the End of Multicore Scaling, in *38th Annual International Symposium on Computer Architecture*, (ACM, San Jose, 2011), pp. 365–376
11. J. Henkel, H. Khdr, S. Pagani, M. Shafique, New Trends in Dark Silicon, in *Proceedings of the 52nd Annual Design Automation Conference*, (ACM, New York, 2015)
12. R. Landauer, Irreversibility and heat generation in the computing process. IBM J. Res. Dev. **5**, 183–191 (1961)
13. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
14. B. Lambson, J. Bokor, Temperature Dependence of Heat Dissipation During Landauer Erasure of Nanomagnets, in *IEEE-NANO, 2012, 12th IEEE Conference on Nanotechnology*, (IEEE, Birmingham, 2012), pp. 1–3
15. L.B. Kish, Thermal (noise) death of Moore's law? AIP Conf. Proc. **665**, 469–476 (2003)
16. P.J.B. Koeck, Quantization errors in averaged digitized data. Signal Process. **81**, 345–356 (2001)
17. K. Roy, S. Bandyopadhyay, J. Atulasimha, Hybrid spintronics and straintronics: a magnetic technology for ultra low energy computing and signal processing. Appl. Phys. Lett. **99**, 063108 (2011)
18. K. Galatsis, A. Khitun, R. Ostroumov, K.L. Wang, W.R. Dichtel, E. Plummer, J.F. Stoddart, J.I. Zink, J.Y. Lee, Y.-H. Xie, K.W. Kim, Alternate state variables for emerging nanoelectronic devices. IEEE Trans. Nanotechnol. **8**, 66–75 (2009)
19. M.S. Fashami, K. Roy, J. Atulasimha, S. Bandyopadhyay, Magnetization dynamics, Bennett clocking and associated energy dissipation in multiferroic logic. Nanotechnology **22**, 155201 (2011)
20. S. Bandyopadhyay, M. Cahay, Electron spin for classical information processing: A brief survey of spin-based logic devices, gates and circuits. Nanotechnology **20**, 412001 (2009)
21. V.V. Zhirnov, R.K. Cavin, J.A. Hutchby, G.I. Bourianoff, Limits to binary logic switch scaling - a Gedanken model. Proc. IEEE **91**, 1934–1939 (2003)
22. S. Salahuddin, S. Datta, Interacting systems for self-correcting low power switching. Appl. Phys. Lett. **90**, 093503–093503 (2007)
23. R.P. Cowburn, D.K. Koltsov, A.O. Adeyeye, M.E. Welland, D.M. Tricker, Single-domain circular nanomagnets. Phys. Rev. Lett. **83**, 1042–1045 (1999)
24. H.G. Gauch, *Scientific Method in Practice* (Cambridge University Press, Cambridge, 2003)

25. E. Fredkin, T. Toffoli, Conservative logic. Int. J. Theor. Phys. **21**, 219–253 (1982)
26. T. Toffoli, *Reversible Computing. Tech. Memo MIT/LCS/TM-151* (MIT Labaratory for Computer Science, Cambridge, 1980)
27. N. Takeuchi, Y. Yamanashi, N. Yoshikawa, Reversible logic gate using adiabatic superconducting devices. Sci. Rep. **4**, 6354 (2014)
28. N. Takeuchi, Y. Yamanashi, N. Yoshikawa, Reversible computing using adiabatic superconductor logic. Lect. Notes Comput. Sci. **8507**, 15–25 (2014)
29. S. Sze, K.K. Ng, *Physics of Semiconductor Devices* (Wiley, Hoboken, 2006)
30. M.W.R. Dreslinski, D. Blaauw, D. Sylvester, T. Mudge, Near-threshold computing: Reclaiming Moore's law through energy efficient integrated circuits. Proc. IEEE **98**, 253–266 (2010)
31. D. Markovic, C.C. Wang, L.P. Alarcon, T.T. Liu, J.M. Rabaey, Ultralow-power design in near-threshold region. Proc. IEEE **98**, 237–252 (2010)
32. H. Soeleman, K. Roy, B. Paul, Robust Ultra-low Power Sub-threshold DTMOS Logic, in *Islped '00: Proceedings of the 2000 International Symposium on Low Power Electronics and Design*, (IEEE, New York, 2000), pp. 25–30
33. S. Fisher, A. Teman, D. Vaysman, A. Gertsman, O. Yadid-Pecht, Ultra-low Power Subthreshold Flip-flop Design, in *Iscas: 2009 IEEE International Symposium on Circuits and Systems*, vol. 1-5, (IEEE, New York, 2009), pp. 1573–1576
34. S. Fisher, A. Teman, D. Vaysman, A. Gertsman, O. Yadid-Pecht, A. Fish, Digital Subthreshold Logic Design - Motivation and Challenges, in *2008 IEEE 25th Convention of Electrical and Electronics Engineers in Israel*, (IEEE, New York, 2008), pp. 682–686
35. V.I. Starosel'skii, Adiabatic logic circuits: A review. Russ. Microelectron. **31**, 37–58 (2001)
36. A. Mishra, N. Singh, Low power circuit design using positive feedback adiabatic logic (PFAL). Int. J. Sci. Res. **3**(6), 02014110 (2014)
37. S. Houri, G. Billiot, M. Belleville, A. Valentian, H. Fanet, Limits of CMOS technology and interest of NEMS relays for adiabatic logic applications. IEEE Trans. Circuits Syst. I Regul. Pap. **62**, 1546–1554 (2015)
38. S. Younis, Asymptotically Zero Energy Computing Using Split-level Charge Recovery Logic. Doctoral Dissertation, Massachusets Institute of Technology, 1994
39. N. Weste, D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspectives*, 4th edn. (Addison-Wesley, Boston, 2010)
40. C. Pawashe, K. Lin, K.J. Kuhn, Scaling limits of electrostatic nanorelays. IEEE Trans. Electron Devices **60**, 2936–2942 (2013)
41. A. Peschot, C. Qian, T.J.K. Liu, Nanoelectromechanical switches for low-power digital computing. Micromachines **6**, 1046–1065 (2015)
42. H.F.G. Pillonnet, S. Houri, Adiabatic Capacitive Logic: A Paradigm for Low-power Logic, in *IEEE International Symposium on Circuits and Systems*, (IEEE, Baltimore, 2017)
43. R.K. Kummamuru, J. Timler, G. Toth, C.S. Lent, R. Ramasubramaniam, A.O. Orlov, G.H. Bernstein, G.L. Snider, Power gain in a quantum-dot cellular automata latch. Appl. Phys. Lett. **81**, 1332–1334 (2002)
44. M.A. Ratner, Intra-molecular electron-transfer - some thoughts and possible applications. Abstr. Pap. Am. Chem. Soc. **174**, 81–81 (1977)
45. M. Ratner, Molecular electronics - pushing electrons around. Nature **404**, 137–138 (2000)
46. C.A. Mirkin, M.A. Ratner, Molecular electronics. Annu. Rev. Phys. Chem. **43**, 719–754 (1992)
47. M. Ratner, A brief history of molecular electronics. Nat. Nanotechnol. **8**, 378–381 (2013)
48. J.P. Bergfield, M.A. Ratner, Forty years of molecular electronics: non-equilibrium heat and charge transport at the nanoscale. Phys. Status Solidi B **250**, 2249–2266 (2013)
49. M.A. Reed, Molecular electronics - back under control. Nat. Mater. **3**, 286–287 (2004)
50. M.A. Reed, Prospects for molecular-scale electronics. MRS Bull. **26**, 113–120 (2001)
51. M.A. Reed, Molecular-scale electronics. Proc. IEEE **87**, 652–658 (1999)
52. J.M. Tour, W.A. Reinerth, L. Jones, T.P. Burgin, C.W. Zhou, C.J. Muller, M.R. Deshpande, M.A. Reed, Recent advances in molecular scale electronics. Ann. N. Y. Acad. Sci. **852**, 197–204 (1998)

53. K. Sotthewes, V. Geskin, R. Heimbuch, A. Kumar, H.J.W. Zandvliet, Molecular electronics: the single-molecule switch and transistor. APL Mater. **2**, 010701 (2014)
54. C.S. Lent, P.D. Tougaw, Bistable saturation in coupled quantum-dot cells. J. Appl. Phys. **74**, 6227–6233 (1993)
55. C.S. Lent, P.D. Tougaw, Lines of interacting quantum-dot cells: a binary wire. J. Appl. Phys. **74**, 6227–6233 (1993)
56. C.S. Lent, P.D. Tougaw, W. Porod, G.H. Bernstein, Quantum cellular automata. Nanotechnology **4**, 49–57 (1993)
57. I. Amlani, A.O. Orlov, G.L. Snider, G.H. Bernstein, Differential charge detection for quantum-dot cellular automata. J. Vac. Sci. Technol. B **15**, 2832–2835 (1997)
58. I. Amlani, A.O. Orlov, G. Toth, G.H. Bernstein, C.S. Lent, G.L. Snider, Digital logic gate using quantum-dot cellular automata. Science **284**, 289–291 (1999)
59. A.O. Orlov, I. Amlani, G.H. Bernstein, C.S. Lent, G.L. Snider, Realization of a functional cell for quantum-dot cellular automata. Science **277**, 928–930 (1997)
60. A.O. Orlov, I. Amlani, R.K. Kummamuru, R. Ramasubramaniam, G. Toth, C.S. Lent, G.H. Bernstein, G.L. Snider, Experimental demonstration of clocked single-electron switching in quantum-dot cellular automata. Appl. Phys. Lett. **77**, 295–297 (2000)
61. C.S. Lent, Molecular electronics – bypassing the transistor paradigm. Science **288**, 1597 (2000)
62. C.S. Lent, B. Isaksen, M. Lieberman, Molecular quantum-dot cellular automata. J. Am. Chem. Soc. **125**, 1056–1063 (2003)
63. C.S. Lent, M. Liu, Y.H. Lu, Bennett clocking of quantum-dot cellular automata and the limits to binary logic scaling. Nanotechnology **17**, 4240–4251 (2006)
64. M.T. Niemier, P.M. Kogge, QCA Circuits, in *Proceedings of the Ninth Great Lakes Syposium on VLSI*, (IEEE, New York, 1999), pp. 118–121
65. M. Mitic, M.C. Cassidy, K.D. Petersson, R.P. Starrett, E. Gauja, R. Brenner, R.G. Clark, A.S. Dzurak, C. Yang, D.N. Jamieson, Demonstration of a silicon-based quantum cellular automata cell. Appl. Phys. Lett. **89**, 0135503 (2006)
66. R.P. Cowburn, M.E. Welland, Room temperature magnetic quantum cellular automata. Science **287**, 1466–1468 (2000)
67. M.B. Haider, J.L. Pitters, G.A. DiLabio, L. Livadaru, J.Y. Mutus, R.A. Wolkow, Controlled coupling and occupation of silicon atomic quantum dots at room temperature. Phys. Rev. Lett. **102**, 046805 (2009)
68. R.D. Brown, J.M. Coman, J.A. Christie, R.P. Forrest, C.S. Lent, S.A. Corcelli, K.W. Henderson, S.A. Kandel, Evolution of metastable clusters into ordered structures for 1,1 '-Ferrocenedicarboxylic acid on the Au(111) surface. J. Phys. Chem. C **121**, 6191–6198 (2017)
69. C.S. Lent, K.W. Henderson, S.A. Kandel, S.A. Corcelli, G.L. Snider, A.O. Orlov, P.M. Kogge, M.T. Niemier, R.C. Brown, J.A. Christie, N.A. Wasio, R.C. Quardokus, R.P. Forrest, J.P. Peterson, A. Silski, D.A. Turner, E.P. Blair, Y.H. Lu, Molecular Cellular Networks: a Non von Neumann Architecture for Molecular Electronics, in *2016 IEEE International Conference on Rebooting Computing (ICRC)*, (IEEE, New York, 2016)
70. J.A. Christie, R.P. Forrest, S.A. Corcelli, N.A. Wasio, R.C. Quardokus, R. Brown, S.A. Kandel, Y.H. Lu, C.S. Lent, K.W. Henderson, Synthesis of a neutral mixed-valence diferrocenyl carborane for molecular quantum-dot cellular automata applications. Angew. Chem. Int. Ed. Engl. **54**, 15448–15451 (2015)
71. N.A. Wasio, R.C. Quardokus, R.D. Brown, R.P. Forrest, C.S. Lent, S.A. Corcelli, J.A. Christie, K.W. Henderson, S.A. Kandel, Cyclic hydrogen bonding in indole carboxylic acid clusters. J. Phys. Chem. C **119**, 21011–21017 (2015)
72. J. Christie, R. Forrest, S. Corcelli, N. Wasio, R. Quardokus, S. Kandel, Y.H. Lu, C. Lent, A. Oliver, K. Henderson, Molecular Switches: Exploring the Counter-ion Problem, in *Abstracts of Papers of the American Chemical Society*, vol. 249, (American Chemical Society, Washington, D.C., 2015)
73. R.C. Quardokus, N.A. Wasio, R.D. Brown, J.A. Christie, K.W. Henderson, R.P. Forrest, C.S. Lent, S.A. Corcelli, S.A. Kandel, Hydrogen-bonded clusters of 1,1′-ferrocenedicarboxylic acid on Au(111) are initially formed in solution. J. Chem. Phys. **142**, 101927 (2015)

74. R.C. Quardokus, N.A. Wasio, J.A. Christie, K.W. Henderson, R.P. Forrest, C.S. Lent, S.A. Corcelli, S.A. Kandel, Hydrogen-bonded clusters of ferrocenecarboxylic acid on Au(111). Chem. Commun. **50**, 10229–10232 (2014)
75. R.C. Quardokus, N.A. Wasio, R.P. Forrest, C.S. Lent, S.A. Corcelli, J.A. Christie, K.W. Henderson, S.A. Kandel, Adsorption of diferrocenylacetylene on Au(111) studied by scanning tunneling microscopy. Phys. Chem. Chem. Phys. **15**, 6973–6981 (2013)
76. H. Qi, S. Sharma, Z. Li, G.L. Snider, A.O. Orlov, C.S. Lent, T.P. Fehlner, Molecular quantum cellular automata cells. Electric field driven switching of a silicon surface bound array of vertically oriented two-dot molecular quantum cellular automata. J. Am. Chem. Soc. **125**, 15250–15259 (2003)
77. H. Qi, A. Gupta, B.C. Noll, G.L. Snider, Y. Lu, C. Lent, T.P. Fehlner, Dependence of field switched ordered arrays of dinuclear mixed-valence complexes on the distance between the redox centers and the size of the counterions. J. Am. Chem. Soc. **127**, 15218–15227 (2005)
78. G. Szakmany, A. Orlov, G. Bernstein, W. Porod, Single-metal nanoscale thermocouples. IEEE Trans. Nanotechnol. **13**, 1234 (2014)

Index

C. S. Lent et al. (eds.), *Energy Limits in Computation*,
https://doi.org/10.1007/978-3-319-93458-7