# A Path Toward Ultra-Low-Energy Computing

Erik P. DeBenedictis*, Michael P. Frank
Center for Computing Research
Sandia National Laboratories
Albuquerque, NM 87185-1319
*epdeben@sandia.gov

Natesh Ganesh, Neal G. Anderson
ECE Department
University of Massachusetts, Amherst
Amherst, MA 01003-9292

***Abstract*—At roughly *kT* energy dissipation per operation, the thermodynamic energy efficiency "limits" of Moore's Law were unimaginably far off in the 1960s. However, current computers operate at only 100-10,000 times this limit, forming an argument that historical rates of efficiency scaling must soon slow. This paper reviews the justification for the ~*kT* per operation limit in the context of processors for von Neumann-class computer architectures of the 1960s. We then reapply the fundamental arguments to contemporary applications and identify a new direction for future computing in which the ultimate efficiency limits would be much further out. New nanodevices with high-level functions that aggregate the functionality of several logic gates and some local memory may be the right building blocks for much more energy efficient execution of emerging applications—such as neural networks.**

***Keywords—logic-memory integration; processing in memory; thermodynamic limits of computing; superconducting circuits***

## I. Introduction

In 1965, Gordon Moore observed that the number of components per integrated circuit was increasing exponentially and predicted that this trend would continue [1]. Together with corresponding increases in the energy efficiency and performance per unit cost of digital logic circuits, this trend enabled exponential growth in the capability, economic utility, and ubiquity of computing systems over the ensuing half-century. However, many observers believe this growth trend will soon slow down or stall due to CMOS approaching physical limits to its energy efficiency [2].

In considering strategies for avoiding this, it is important to distinguish between processing and memory functions. Nonvolatile memory technologies (*e. g.* flash memory) require no power to simply retain stored data, so simply stacking up more layers of memory on a chip will be able to raise the effective areal density of digital storage for some time to come. Moreover, as storage sizes continue to increase, one can co-locate a proportional amount of processing circuitry for an almost negligible extra cost—as long as most of this circuitry is turned off (*i. e.*, not dissipating any power) most of the time. When some local transformation of data is needed, it can happen locally, minimizing the energy cost incurred for data movement in contrast to the traditional approach of a von Neumann computer where the overall system is divided into separate processing and memory subsystems with a long path between them that must be used every time data is accessed.

Provided that integration of logic and memory can minimize energy dissipation for data movement, the problem of how to minimize the energy dissipation for the logic itself remains. Landauer [3] observed that there is a fundamental thermodynamic limit of energy dissipation for logically irreversible operations (those that cause a merging of digital states) of a magnitude that is proportional to the reduction in Shannon entropy of the digital state ensemble. For the class of "typical" operations that Landauer studied in detail, namely, traditional Boolean logic operations with unknown (and equiprobable) inputs that are not preserved, the minimum dissipation is on the order of $kT$, where $k$ is Boltzmann's constant and $T$ is the temperature of the system's thermal environment. In the case of the irreversible erasure of exactly one bit of information that is equally likely to have been in the 0 or 1 state before erasure, the limit comes out to $kT \ln 2$. This formula is frequently cited as constituting a general limit on energy dissipation for digital logic operations, but this can be misleading for two reasons:

First, the exact magnitude of the Landauer limit depends on the type of logic operation being considered. For example, reversible operations do not reduce the entropy of the digital state ensemble at all, and theoretically do not require any minimum energy dissipation; yet they are still computationally universal [4]. Unfortunately, pure reversible computations generally incur some algorithmic overheads [5].

Second, even in the case of operations that are not perfectly reversible, the exact magnitude of the Landauer limit depends on the probabilities that states will be merged and thus also on the relative probabilities of the various inputs. These considerations should be taken into account when considering the Landauer limit in new contexts.

It is often argued that these fundamental thermodynamic limits are not practically relevant, because the energy efficiency of logic will plateau long before the fundamental limits are reached unless formidable practical challenges are met. However, if these challenges *are* successfully met, and efficiency continues scaling at near historical rates, a gap of 100-10,000× between fundamental and practical limits will close within a few decades.

Approved for unclassified unlimited release SAND2016-7365 C. Research supported by Sandia National Laboratories, a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the US Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

978-1-5090-1370-8/16/$31.00 ©2016 IEEE

In this paper, we identify a class of conceptually useful digital operations that are irreversible—and thus can avoid the overhead of fully-reversible logic—but for which the Landauer minimum energy cost can be well below $kT$ ln 2 per operation under inputs characteristic of contemporary applications. The example we provide—the updating of a digital "synapse" when presented with data to be learned—is relevant in the context of brain-inspired "neuromorphic" computing schemes that are currently under consideration.

## II. Clarifying the Landauer Limit

We begin by revisiting the reasoning from Landauer's 1961 paper [3] that led to what is now widely known as the "Landauer limit"—although the current meaning of that phrase is not from the paper. Landauer considered a physical device in contact with an environment at temperature $T$, and a finite number of distinguishable states of this device that are used to encode data. He associated the quantity

$$S = -k\sum_j p_j \log_e p_j \qquad (1)$$

with the contribution of the information-bearing degrees of freedom to the thermodynamic entropy of the device, where $j$ labels the device data states and $p_j$ denotes their respective probabilities of occurrence. Now, if the device undergoes a transformation that deterministically maps initial data states having nonzero probabilities into a smaller number of final data states, then the final entropy $S_f$ is necessarily smaller than the initial entropy $S_i$. Thermodynamically, Landauer argued, this requires that the entropy of the surrounding thermal environment increase by at least an amount $S_i - S_f$, which in turn requires an environmental heating of at least $(S_i - S_f)T > 0$. This is Landauer's Principle. It has become customary to express this relation in terms of the Shannon entropy (or Shannon "information") of the data-state probability distribution, expressed in units of "bits" as

$$H = -\sum_j p_j \log_2 p_j. \qquad (2)$$

With this, the environmental heating is

$$\Delta E_{env} \geq (kT \ln 2)(H_i - H_f). \qquad (3)$$

This is origin of the "Landauer limit," as it is most commonly known. It specifies a lower bound on the dissipative cost of "$kT$ ln 2 per lost bit" in logically irreversible (many-to-few) transformations, and specifically as "$kT$ ln 2 per erased bit" for erasure (many-to-one) transformations that map all initial states into a single final state (so $H_f = 0$). (It should be noted that "bit" is used here as a unit of information, and may be fractional.)

**A. Landauer's analysis of AND gate and wire (figure 5 from [3])**

| Si terms | | Before Operation | | | | After Operation | | | Sf terms | |
|---|---|---|---|---|---|---|---|---|---|---|
| Prob. | (in k's) | p | q | r | | p1 | q1 | r1 | Prob. | (in k's) |
| 0.1250 | 0.2599 | 1 | 1 | 1 | → | 1 | 1 | 1 | 0.1250 | 0.2599 |
| 0.1250 | 0.2599 | 1 | 1 | 0 | → | 0 | 0 | 1 | 0.1250 | 0.2599 |
| 0.1250 | 0.2599 | 1 | 0 | 1 | → | 1 | 1 | 0 | 0.3750 | 0.3678 |
| 0.1250 | 0.2599 | 1 | 0 | 0 | → | 0 | 0 | 0 | 0.3750 | 0.3678 |
| 0.1250 | 0.2599 | 0 | 1 | 1 | → | 1 | 1 | 0 | | |
| 0.1250 | 0.2599 | 0 | 1 | 0 | → | 0 | 0 | 0 | | |
| 0.1250 | 0.2599 | 0 | 0 | 1 | → | 1 | 1 | 0 | | |
| 0.1250 | 0.2599 | 0 | 0 | 0 | → | 0 | 0 | 0 | | |
| Si: | **2.0794** | **k** | | | | | | | Sf (k's) | **1.2555** |

**B. Alternative presentation of Landauer's table:** Si - Sf (k's) **0.8240**

| | | p | q | r | | p1 | q1 | r1 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.1250 | 0.2599 | 0 | 0 | 0 | → | 0 | 0 | 0 | 0.3750 | 0.3678 |
| 0.1250 | 0.2599 | 0 | 0 | 1 | → | 0 | 0 | 1 | 0.1250 | 0.2599 |
| 0.1250 | 0.2599 | 0 | 1 | 0 | → | 1 | 1 | 0 | 0.3750 | 0.3678 |
| 0.1250 | 0.2599 | 0 | 1 | 1 | → | 1 | 1 | 1 | 0.1250 | 0.2599 |
| 0.1250 | 0.2599 | 1 | 0 | 0 | | | | | Sf (k's): | **1.2555** |
| 0.1250 | 0.2599 | 1 | 0 | 1 | | | | | Si - Sf (k's): | **0.8240** |
| 0.1250 | 0.2599 | 1 | 1 | 0 | | | | | | |
| 0.1250 | 0.2599 | 1 | 1 | 1 | | | | | | |
| Si: | **2.0794** | **k** | | | | | | | | |

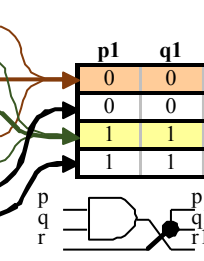



Fig. 1. Example from [3].

For illustrative purposes, we consider evaluation of the Landauer "limit" in detail for a specific example—one from Landauer's original paper [3]—that involves a common Boolean operation (AND). The truth table, rendered in the visually distinctive white characters and black background of [3], is shown for in Fig. 1A for this circuit (diagram in Fig. 1B inset). The truth table is represented in a new form in Fig. 1B that better highlights the initial-to-final state mergings that occur in this logically irreversible transformation. Probabilities $p_j$, the entropies $S_i$, $S_f$, and their difference $(S_i - S_f)$ (in units of $k$) are all tabulated for the case where the eight possible input vectors are equiprobable—also as assumed by Landauer. The entropy change is $S_i - S_f = 0.824\ k \approx k$, corresponding to a lower bound on the energy dissipation of $\sim kT$.[1]

We emphasize that, as is clear in Fig. 1, the value of $\sim kT$ energy dissipation per use obtained for this example is as much a result of the assumed input probability distribution as it is of the state mapping implemented by the gate. Some groups of input states merge into a single output state whose probability is equal to the sum of the probabilities of the contributing input states: the three input combinations $pq$=00, 01, and 10 to an AND gate merge into the single output state $r_1$=0, which lowers the entropy, whereas only the $pq$=11 input state yields the $r_1$=1 output state, and so does not contribute to the entropy change. Thus, we can say that although the particular transformation in Fig. 1 is not fully logically reversible, it is partially or *conditionally* reversible (a notion elaborated upon in [7]); that is, there is a certain precondition on the inputs (here $pq$=11) under which no state mergings will occur. In the general case, the entropy reduction associated with the full transformation depends on the input probabilities, which

[1] Landauer's original paper miscalculated the entropy difference of this example as 1.18 $k$. This was later corrected in [6].

"prescale" contributions from all of the various inputs—those that satisfy the precondition as well as those that do not.

The assumption of a uniform input distribution, almost ubiquitous since Landauer, is entirely reasonable for common Boolean gates and logic circuits operating as they might in an unspecified general-purpose machine executing an unknown computational task. Uniform probabilities are assigned when there is no reason to expect otherwise. Under this assumption, the information loss for most common Boolean logic gates is $H_i - H_f \sim 1$ bit *per use*, yielding a Landauer limit of $\sim kT$ energy dissipation per use for uniformly distributed inputs. Since Boolean gates have been the established building blocks of digital computers for over half a century, and since the ubiquitous assumption of uniform probabilities has seemed reasonable for gates and logic circuits in the kinds of digital computers that have been in use during this period, the Landauer limit is often interpreted as a dissipation bound of at least "$kT$ ln 2 per use" (or "$kT$ ln 2 per operation"). This is a useful shorthand under the assumptions that justify it, but only under these assumptions.

In cases like those of interest in this paper, where input probabilities are expected to be highly skewed, Landauer's original argument must be revisited if it is to be properly applied. In such cases, evaluation of the Landauer minimum can yield dissipation bounds much lower than $kT$ ln 2. This obviously conflicts with the "shorthand" Landauer limit of $kT$ ln 2 energy dissipation per use, but not with the "actual" Landauer minumum of $kT$ ln 2 per lost bit calculated as above for nonuniformly distributed inputs. There is no contradiction—far less than one bit per use can be lost on average when the input distribution is highly skewed and the information loss is $H_i - H_f < H_i \ll 1$ bit.

Nonuniform input distributions can thus yield dissipation bounds lower than $kT$ ln 2 per use with no violation of the Landauer limit as defined above. We should emphasize that although some have questioned Landauer's assumptions and his application of equilibrium thermodynamics to this problem, his essential result—a dissipative contribution of $kT$ ln 2 per bit of irreversible information loss—is upheld in a wide variety of proofs and derivations that sidestep these objections and even quantify information differently (*e.g.* [8], [9], [10]). We should also note the distinction between information loss reductions resulting from skewed input distributions, which reduce the probability of state mergers overall, and elimination of information loss by eliminating state merging altogether as in reversible computing [4]. Finally, we note that acceptance of the Landauer limit does not amount to a claim that it can be achieved. We discuss both Landauer limit reductions in scenarios with heavily skewed input distributions and the achievability of these reduced limits in the following sections.

## III. A Simple Learning Machine

We now apply the analysis of [3] to a device with a functionality and input environment inspired by emerging applications such as neuromorphic computing. Instead of an AND gate with uniform inputs, we will consider an artificial synapse of sorts with a nonuniform input distribution and show that the minimum energy dissipation per operation can be much less than $kT$.

While learning is essential, most experiences do not cause a given synapse to change state. We will exploit the low probability of actual learning to lower minimum energy. For example, readers of this paper will have already learned the alphabet as a child. By now, there is nothing more to learn by seeing the letter "L" one more time. However, seeing the letter "Л" may invoke learning and cause synapse changes for readers who are unfamiliar with the letter equivalent to "L" in Russian (Cyrillic). This will be a rare event.

We consider a single simplified artificial synapse as the machine in our example, and analyze a system comprising an array of these machines. The system is a functionally enhanced memory tasked with learning or creating a model of a slowly changing environment from partial observations. The environment comprises of an array of $n \times n$ (here $n$=3) data items or pixels that take the values −1 and +1. We will ultimately analyze two different scenarios for the environment, one where all the pixels are spatially independent and the other where the pixels in a row are perfectly correlated. Observations are of one pixel (or row) at a time, with probability $p$ that a specific pixel (respectively, row) is observed in each step in cases of spatially independent (respectively, correlated) pixels. The system has an internal $n \times n$ array of functionally enhanced storage cells and shift registers that drives both the row and column of the internal array with the observed pixel value of −1 or +1. When the selected cell receives (−1, −1) or (+1, +1), it remembers the stimulus value. Each pixel in the environment changes with time at a rate corresponding to a probability $q$ of a change per observation. The system will be modeled in steady state, so an initial condition is not needed. Table I is an example data set corresponding to the problem description above. The system could drive multiple rows and columns at once and include both −1 and +1 data values in the same observation, but this will not be considered here.

An implementation of the example system is illustrated in Fig. 2A, which is an $n \times n$ array of the synapse machines in a framework that transmits data in Table I past the array as shown. The function being analyzed will be just one of the synapses in the array, which is modeled as a magnetic core. Magnetic cores are used as a behavioral illustration at this point

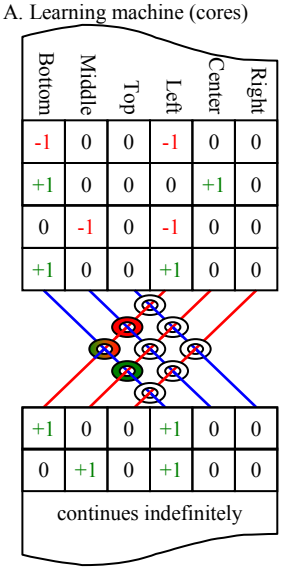
TABLE I: Data to be Learned

| *Step* | *Row* | *Column* | *Response* |
|---|---|---|---|
| 1 | −1 on bottom | −1 on left | Learn −1 |
| 2 | −1 on middle | −1 on left | Learn −1 |
| 3 | +1 on bottom | +1 on center | Learn +1 |
| 4 | −1 on bottom | −1 on left | −1 already learned |
| many repetitions with no learning | | | |
| $n$-3 | −1 on bottom | −1 on left | −1 already learned |
| $n$-2 | −1 on middle | −1 on left | −1 already learned |
| $n$-1 | +1 on bottom | +1 on center | +1 already learned |
| $n$ | +1 on bottom | +1 on left | Learn +1 |

because readers are likely to be familiar with their operation, but we will mention a nanodevice (MeRAM) before the end of this section that is compatible with the same analysis.

The system monitors a stream of $2n$ parallel data inputs from the environment (one for each row and column), which is assumed to be ongoing and which is not destroyed or erased by the system. For the case of single pixel observations, the stream provides a single nonzero, ±1, stimulus on each set of $2n$ data inputs as shown in Fig. 2A to write into the corresponding core. (In the case of the spatially correlated environment, the stream contains multiple ±1 inputs to update an entire row of cores with the same value.) As the data flows downward through the $2n$ shift registers, the values on the bottom row are translated into current in the blue and red wires. The wires become rows and columns of an array tilted at 45° where the row-column intersections each flow through the center of a core. Each core flips to align with its magnetic field, but only if the field is above a threshold and a core will not flip if it is already in the correct state. The system would be engineered to flip magnetization at ±1.5 units of current flowing through each core. Thus, a core exposed to +1 on the row wire and +1 on the column wire will have total current +2 and would flip magnetization to the green state provided it was not in the right state already. Vice versa for −1 and a red state. Magnetic cores dissipate energy when they change state, but nearly zero energy otherwise. Unless the two currents are in the same direction, the total current will be below the threshold and there will no state change and no energy dissipation associated with core state changes.

A. Learning machine (cores)




B. MeRAM equiv.

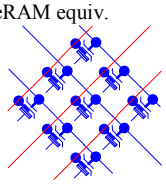


Fig. 2: Two versions of system

Fig. 2A illustrates the system processing the data in Table I, specifically at the processing of step $n$. Steps 1-3 cause the system to learn pixels, setting the three non-white cores shown in Fig. 2A; the white cores are irrelevant to the discussion and could be either red or green. The system then experiences a long sequence of steps containing repeating known pixels. In the last row of Table I, the learning machine observes a change in the external data set. The {bottom, left} pixel changes from −1 to 1 and is recorded as the leftmost core in Fig. 2A flips. We now consider lower bounds on the energy dissipation for this machine.

## IV. Dissipation Analysis for the Learning Machine

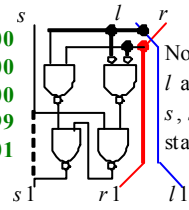
In this section, we obtain lower dissipation bounds for the learning machine of Sec. III, Each magnetic core behaves as a finite-state automaton, as does the entire learning machine. We consider both of the scenarios for the pixel environment and the input streams mentioned in the previous section. We will start with a limiting dissipation analysis of a single core, which will apply equally to both cases. We will then calculate the limiting dissipation of the entire learning machine and elucidate the differences in the dissipation for the two cases.

Dissipation bounds are obtained from a fundamental physical description of Finite State Automata (FSA) driven by Independent Identically Distributed (IID) information sources [11], extended for the present paper to accommodate FSA driven by inputs with temporal correlations and thus for learning scenarios in changing environments. Landauer's focus was combinational logic, but his analysis can be applied a manner that yields the same result for the case at hand (see Fig. 3).

The FSA description of each core is as follows: The FSA state $s$ corresponds to the current magnetization state of the core. FSA inputs $l$ and $r$ correspond to the current states in the blue and red wire respectively. The next state of the core $s'$ depends upon its current state $s$ and the input values on the wires. We use the random variables $S$, $S'$, $L$ and $R$ for a statistical description of the current and next state of the core, and for the two inputs, respectively. Assuming that the magnetization states of the core are perfectly distinguishable, the minimum energy dissipated into the environment as the

**Learning Machine (Synapse)**

| | |
|---|---|
| Probability of seeing learnable data (+,+ or -,- | 0.0100 |
| Probability data has changed since last learne | 0.0100 |
| P(null) | 0.9900 |
| P(reinforce) | 0.0099 |
| P(new data) | 0.0001 |




Notes: $r$ and $l$ are trits & $s$, $s$ 1 are state

| Prob. | Si terms (in k's) | left | right | field | left | right | field | Prob. | Sf terms in (k's) |
|---|---|---|---|---|---|---|---|---|---|
| 0.00495 | 0.0263 | -1 | -1 | -1 | -1 | -1 | -1 | 0.0050 | 0.0265 |
| 0.09900 | 0.2290 | -1 | 0 | -1 | -1 | 0 | -1 | 0.0990 | 0.2290 |
| 0.09900 | 0.2290 | 0 | -1 | -1 | 0 | -1 | -1 | 0.0990 | 0.2290 |
| 0.09900 | 0.2290 | 0 | 0 | -1 | 0 | 0 | -1 | 0.0990 | 0.2290 |
| 0.09900 | 0.2290 | 0 | 1 | -1 | 0 | 1 | -1 | 0.0990 | 0.2290 |
| 0.09900 | 0.2290 | 1 | 0 | -1 | 1 | 0 | -1 | 0.0990 | 0.2290 |
| 0.00005 | 0.0005 | 1 | 1 | -1 | -1 | 0 | 1 | 0.0990 | 0.2290 |
| 0.00005 | 0.0005 | -1 | -1 | 1 | 0 | -1 | 1 | 0.0990 | 0.2290 |
| 0.09900 | 0.2290 | -1 | 0 | 1 | 0 | 0 | 1 | 0.0990 | 0.2290 |
| 0.09900 | 0.2290 | 0 | -1 | 1 | 0 | 1 | 1 | 0.0990 | 0.2290 |
| 0.09900 | 0.2290 | 0 | 0 | 1 | 1 | 0 | 1 | 0.0990 | 0.2290 |
| 0.09900 | 0.2290 | 0 | 1 | 1 | 1 | 1 | 1 | 0.0050 | 0.0265 |
| 0.09900 | 0.2290 | 1 | 0 | 1 | | | | Sf (k's): | **2.3425** |
| 0.00495 | 0.0263 | 1 | 1 | 1 | | | | **Si - Sf (k's):** | **0.0006** |
| Si: | **2.3431** | **k** | | | | | | | |

Fig. 3. Analysis method of [Landauer 61] applied to synapse function

core (in steady state) undergoes a transition from $s$ to $s'$ is $\Delta E_{env} \geq kT \ln 2\ [H(S|LR) - H(S'|LR)]$ per operation where $H(S|LR)$ and $H(S'|LR)$ are the conditional Shannon entropies of the core state distribution given the inputs, before and after the state transition respectively. The inputs $(l, r) = (+1, +1)$ and $(l, r) = (-1, -1)$ write +1 and −1 into the core states respectively, regardless of the previous state. This merging of the core states for certain $l$ and $r$ inputs is the source of the irreversibility and energy dissipation into the environment.

We have calculated the limiting dissipation for the learning machine with $p = 0.01$ and $q = 0.01$. Recall that $p$ is the probability of seeing learnable data, i.e. the probability of seeing the inputs $(l, r) = (+1, +1)$ or $(l, r) = (-1, -1)$. $q$ is the probability that given the presence of learnable data, the data value changes in the environment since the last time that data was observed. The input probabilities are functions of $p$, $q$, and the steady state core state distribution is $P(S=+1) = P(S=-1) = 0.5$. The lower bound on energy dissipation calculated for a single core of the learning machine—both from the FSA description and the modified Landauer-like analysis of Fig. 3—is $\Delta E_{env} \geq 0.0006\ kT$ per operation. The 1,000× difference between the limiting dissipation for the magnetic core and the "$kT$ ln 2 per operation" rule of thumb stems largely from the input probabilities selected for this learning example, which correspond to learning with a slowly-changing environment.

We now extend our analysis to the entire learning machine for the two scenarios introduced in the previous section. The magnetic cores are assumed not to interact with one other. In the first case, the pixels in the 3×3 environment are spatially independent and the cores updated one at a time randomly. The limiting dissipation bound for the entire learning machine will be equal to the sum of the dissipation bounds for the nine individual cores. For $p = 0.01$ and $q = 0.01$, we have the lower bound on the energy dissipated into the environment for the nine-core learning machine to be $\Delta E_{env} \geq 9 \times 0.0006\ kT = 0.0054\ kT$. In the second case, updating an entire row with correlated inputs, will produce correlations between the cores of each row. As a result, the limiting dissipation of the entire learning machine will be < 9 times that of a single core. Using the same values for $p$ and $q$ as before, we have the lower bound on the energy dissipation of the learning machine of $\Delta E_{env} > 0.00168\ kT$. Thus, the limiting dissipation values for variations of the learning machine can vary significantly, depending upon the characteristics of the input environment and the updating scheme employed, even for a fixed limiting dissipation values for the individual cores.

We next consider the principle of aggregation, which we will define as follows: The minimum energy dissipation of a function will always be less than or equal to the minimum for a realization as a disaggregated group of lower level (often non-optimal) primitives. To illustrate, consider the magnetic core from the learning machine. Each of the nine magnetic cores is functionally equivalent to the logic circuit in Fig. 3 comprised of NAND primitives (two of which use three-valued inputs). A dissipation analysis of this circuit using the same input distribution as the magnetic core implementation, and assuming that the gate operations are not conditioned upon $l$ and $r$ inputs, gives a dissipation bound of $\Delta E_{env} > 2.8939\ kT$. This is $\gg 0.0006\ kT$, the large difference attributable to a highly non-optimal disaggregation of the logic function using gate-level primitives. This dramatically illustrates both the aggregation principle and the need for careful analysis and interpretation.

We reiterate that the behavior of a magnetic core is well known to engineers due to its historical use in computers, and thus serves as a suitable example device to illustrate aggregation. However, legacy core memory cells are macroscopic devices and their practical dissipation would be orders of magnitude above the dissipation limits obtained here, both because of dissipation associated with changes in the core magnetization and that associated with generation of the required wire currents on each use.

A MagnetoElectric RAM (MeRAM) [12] is a modern nanodevice that exhibits similar behavior, but with currents replaced by voltages. The MeRAM equivalent of the learning machine's array is shown in Fig. 2B, using the MeRAM schematic symbol and limiting the diagram to cell writing (the device terminal needed for reading is not connected). An MeRAM-based implementation of our synapse would provide an aggregated realization of the function and have a much lower dissipation than a macroscopic core, while still being over the theoretical minimum. We will see how minimal dissipation might actually be approached in an already-available technology in the following section.

## V. Approaching Fundamental Limits

The Landauer limit is a lower bound on dissipation per operation that will be approachable to varying degrees in various technology contexts. We considered an example above, inspired by learning applications, for which the Landauer limit evaluated for individual devices is $\ll kT$ per operation. We now consider the physical possibility of approaching this limit.

As an example of a device that could very nearly achieve the Landauer limit in the learning machine discussed above, we propose and partially analyze a Josephson Junction- (JJ-) based negative-inductance Superconducting QUantum Interference Device (nSQUID) circuit with the behavior needed for the minimum energy model to apply. In contrast to the MeRAM, this circuit appears to have the necessary properties to approach the energy minimum in Fig 3. Furthermore, the key nSQUID subcircuit has been constructed and measured in other contexts (*i.e.,* a shift register, not an array). The measurements show about 1 $kT$ per operation, which is extraordinary by most standards yet above the sub-$kT$ minimum suggested by Fig. 3.

We provide an introduction to the nSQUID, but readers will need to reference [13] for enough details to duplicate the results. The nSQUID circuit illustrated in Fig. 4A has current from the $V_{dc}$ supply pass through the two branches on the left to ground. A common mode bias current $I_+$ originates with $V_{dc}$ and flows through $L_1$ and $L_2$ in the same direction. Current can also flow or circulate in opposite directions through $L_1$ and $L_2$, which we designate $I_-$. Circulating current represents a 0 or 1 data value depending on whether the rotation is clockwise or counterclockwise.

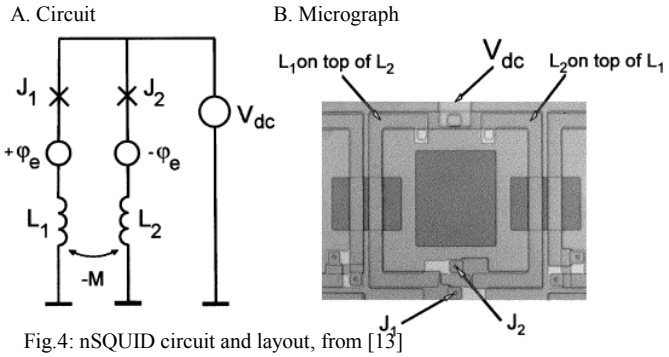


Fig.4: nSQUID circuit and layout, from [13]

The circuit is laid out with $L_1$ and $L_2$ as one-turn inductors wrapping in opposite directions around the empty square in the center of Fig. 4B. Due to the reversed wrapping, $I_+$ flowing equally through $L_1$ and $L_2$ creates no net magnetic field, but magnetic fields from the $I_-$ current representing data adds and creates a larger magnetic field.

Quantum mechanics forces the magnetic field threading a superconducting loop to be quantized, which impacts the circulating current defining data bits, but has no effect on the bias current because there is no magnetic field.

Due to both the effects of quantized magnetic field and the classical inductance, varying the bias current smoothly shifts the circuit from having a single potential to two potentials. Fig. 5 is a plot of the energy in the nSQUID as a function of the current that defines the data $I_-$. The curves vary by the amount of common mode current $I_+$, which rises from low values at the top to higher values as the curves move downward (however, further increase in $I_+$ does not result in a deeper double-welled potential). The units are not relevant to the point of this paper but are the same as in [13].

A key step toward reaching the low energy limit is to properly implement a protocol for erasing information when there is an unequal distribution of 0's and 1's. Three increasingly sophisticated erasure protocols will be described below, with the last being sufficient for the purposes of this paper.

Slowly lowering the energy barrier between data states 0 and 1 is sufficient to achieve dissipation of $kT \ln 2$, which is the minimum possible when $p_0 = p_1 = 0.5$, where $p_j$ is the probability of a bit assuming value $j$.

When $p_0 \neq p_1$, entropy $S$ is less than one bit, and it ought to be possible to erase the information with just $-T\Delta S$ heat generation. The protocol [14] is easy to explain and understand, but achieves optimal efficiency only in the limit of infinite time. Starting with a large energy barrier separating 0 and 1, the first step is to tilt the energy landscape so that the less probable bit value is at a higher energy

$$\Delta E = kT \ln \frac{1-p}{p} \cong kT \ln p^{-1}, \tag{4}$$

where the approximation here approaches equality for small $p$; this value of $\Delta E$ gives an equilibrium high-energy state occupancy probability of $p$ as per the Boltzmann distribution for a two-state system,

$$p = \frac{1}{1+e^{\Delta E/kT}} \cong e^{-\Delta E/kT}, \tag{5}$$

where here the approximation holds for large $\Delta E$. This tilt puts the system into a thermodynamic equilibrium, yet with a high energy barrier that prevents rapid transitions. The barrier is then gradually lowered, which gradually causes the states to merge. This protocol approaches minimum dissipation as the rate of lowering becomes infinitely slow.

The protocol most applicable to this paper erases a nonuniform bit (where $p_0 \neq p_1$) in a specified time $t_f$ with minimum heat generation given that constraint. The required protocol was derived by Zulkowski in [15] and dissipates heat of $-T\Delta S + O(1/t_f)$. It should not be surprising that this exceeds heat dissipation of the previous protocol, but is still minimal. The first term is the cost of erasing the information and the second term is the familiar result that the energy efficiency of adiabatic systems varies inversely with the speed of operation. As described in in [15], the protocol uses waveform $V_t$ for tilt and $V_b$ for the height of the separating barrier. The variable $V$ is usually reserved for voltage, which will be confusing in subsequent discussion because nSQUID circuits are controlled by currents. Therefore we will call the waveforms in [15] $I_t$ and $I_b$. The exact waveforms depend on $p_0$, $p_1$, and the available time $t_f$, but essentially the barrier goes down then up while the tilt goes up.

An nSQUID circuit has essentially the same tilt and barrier height controls as the bit erasure protocol in [15]. Specifically, barrier height is controlled by the current $I_+$ flowing through the circuit in common mode, as illustrated in Fig. 4A. Tilt can be in the form of a magnetic field covering the entire array, or two additional wires shown in Fig. 4A as $+\varphi_e$ and $-\varphi_e$.

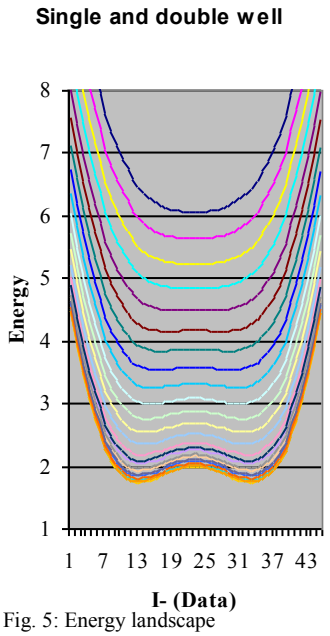


Fig. 5: Energy landscape

However, the example in Fig. 2 also has an array structure with row-column addressing. In general, row-column addressing means each cell receives separate signals from the row and column to which it is connected. These signals arrive in four combinations: unselected, half-selected (in two versions corresponding to just the row or just the column), and selected. While the nSQUID has two controls (tilt and barrier

height), these are not combined properly for row-column addressing. The sum of the row and column currents will control the barrier height of the nSQUID in this paper. In a minor variation of the problem definition, the tilt signal could be applied globally—such as with a magnetic field enveloping the entire array or a single current routed to every cell.

We devise a signaling protocol for both addressing and erasure [15]. In simple words, unselected cells hold data indefinitely. Selection occurs when the erasure waveforms are applied to both the row and column conductors of a cell, in which case they combine and efficiently erase the information. Half-selected cells only get half the erasure waveform, which by careful engineering causes the cell to retain data.

The protocol requires three common mode current levels $I_{select}$, $I_{half}$, and $I_h$ ($I_h$ stands for $I$-hold and is the current when unselected) for the nSQUID such that:

A. The currents are equally spaced and in a particular order, specifically $I_h = I_{half} + \Delta I = I_{select} + 2\Delta I$ for a current spacing $\Delta I$.

B. The nSQUID holds data reliably when $I_{half} \le I_+ \le I_h$, even when the energy landscape is tilted to the maximum required by the protocol in [15].

C. The protocol in [15] will function properly when $I_{select} \le I_b(t) \le I_{half}$, meaning the bit erasure protocol does not require currents outside the range between half-selection and selection. This implies $I_b(t) - I_h \le I_{half} - I_h = -\Delta I$.

Table II shows a way to combine array addressing and erasure; it is laid out like a 2×2 memory with the lower right cell selected. Unselected rows receive no current and columns receive $I_h$, thus causing all unselected cells to hold their state. To select a cell, the cell's row and column each receive a (negative) current change of $\frac{1}{2}(I_b(t)-I_h)$, resulting in the selected cell being exposed to the proper waveform $I_b(t)$ for the erasure protocol. All half-selected cells hold data reliably because they receive a current greater than $I_{half}$.

TABLE II: Currents applied to nSQUID array

| | Unselected column $I_{col} = I_h$ | Selected column $I_{col} = I_h + \frac{1}{2}(I_b(t)-I_h)$ |
|---|---|---|
| Unselected row $I_{row}=0$ | $I_+ = I_h$ | $I_+ = I_h + \frac{1}{2}(I_b(t)-I_h) \le I_{half} + \frac{1}{2}\Delta I$ |
| Selected row $I_{row} = \frac{1}{2}(I_b(t)-I_h)$ | $I_+ = I_h + \frac{1}{2}(I_b(t)-I_h) \le I_{half} + \frac{1}{2}\Delta I$ | $I_+ = I_h - (I_b(t)-I_h) = I_b(t)$ |

The effectiveness of the protocol requires the nSQUID meet requirements A-C. To show feasibility, Fig. 6 includes curves from the nSQUID circuit equations in [13] at an operating point that supports addressing. For addressing, we choose current values of $I_h = 2.2$, $I_{half} = 2.8$, and $I_{select} = 3.4$, which have equal spacing $\Delta I = 0.4$. Fig. 6A shows three curves from Fig. 5 with the values specified above, plus tilt. Two of the curves are bistable and the third is not. For additional assurance, Fig. 6B shows a series of curves $I_{select} \le I_+ \le I_{half}$ where the bistable well decreases in depth.

## VI. Towards a Roadmap for Sub $kT$ Computing

The ideas above include ingredients for the design of new kinds of computing systems with extremely low energy dissipation. While the best CMOS today dissipates about $10^4$ $kT$ per operation, the record for low-loss logic is $E_r \approx 1$ $kT$ [13]. It is reasonable to expect $E_r$ will be reduced to 0.1 $kT$, 0.01 $kT$, and so forth—and similarly for non-logic functions like the synapse example presented earlier. Let us outline steps for the development of ultra low energy computing based on ideas in the preceding sections:

A. To approach the thermodynamic limits of standard Boolean gates in the traditional computing paradigm, it is reasonable to assume equiprobable inputs and irreversible loss of input information, leading to a "rule of thumb" lower dissipation bound of "$kT$ per operation." For such scenarios, this rule of thumb accurately reflects the spirit of Landauer's analysis of [3], but in other scenarios the dissipation bounds must be revisited.

B. Reversible logic styles in the sense of [4], [5], [7] may become viable in the near future. While these can have arbitrarily low dissipation in principle, any specific implementation technology will have some practical minimum dissipation $E_r$ per operation. With $E_r \approx 1$ $kT$ today [13], reversible logic is near the threshold of yielding benefit over conventional logic for some applications.

C. A hybrid of steps A and B could lead to complete systems. Reversible logic creates intermediate variables that must be preserved until they can be decomputed, incurring a cost of ~$E_r$ every time the temporary variable propagates through a reversible gate. Reversible gates from step B could be used when such signals need to propagate $kT/E_r$ steps, otherwise Boolean gates from step A would be used.

D. Minimum energy requirements may be reduced in

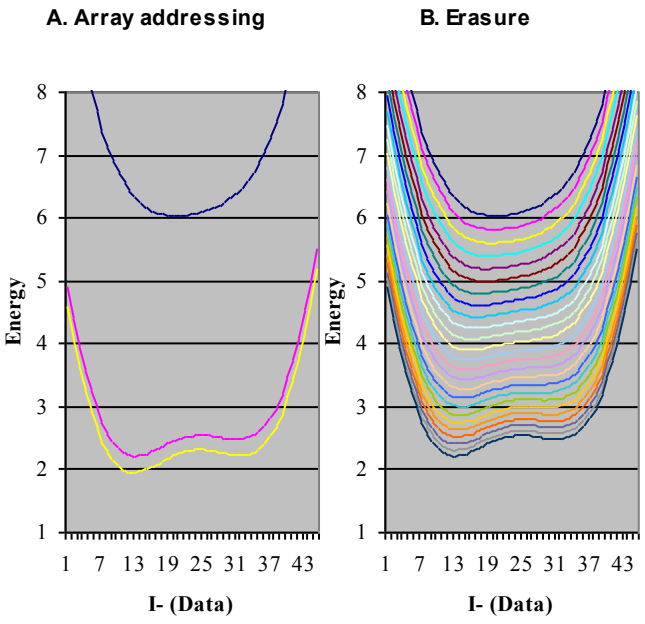



Fig 6: Array addressing and erasure protocol at the same time

contexts where the probabilities of various inputs are nonuniform and known. The options below become available once input probabilities for gates have been found by analysis or simulation.

E. Based on section VI of this paper, the engineer could use a technology-limited Zulkowski eraser as a primitive. While the discussion in section VI discussed asymptotically efficient erasure, let us assume that a real Zulkowski eraser would have parasitic dissipation of $\sim E_r$ because it uses the same technology as reversible logic. This changes criteria C above by making it more effective to erase a signal containing between $\sim E_r$ and $\sim kT/E_r$ information instead of saving and decomputing or erasing it inefficiently.

F. More energy efficient versions of the gates in step A can be designed with advance knowledge of their input distributions from step D. This leads to a general class of thermodynamically-optimized logical primitives, namely, operations that are conditionally reversible [7] (*i.e.*, transform some subset of the input states reversibly). This approach could reach the thermodynamic minimum dissipation for a logic circuit specified in advance, but will not help design the logic circuit in the first place.

G. As an independent research path, the strategies above add motivation for the development of non-von Neumann computer architectures. Gates in a well-designed CPU of a von Neumann computer should have nearly equiprobable input combinations. If not, many gates will be inefficiently used and the design could be improved irrespective of any arguments in this paper. However, it is not bad design for a state-containing device to be idle most of the time because it is serving the useful function of holding information. Therefore, an integrated logic-memory architecture could offer more opportunities to apply items A-F above and thereby reduce dissipation.

H. For all the above steps, discovery of new computing devices could improve energy efficiency through the aggregation principle discussed in Sec. IV. The opportunity is to seek out new electronic devices that perform more and more sophisticated functions. For example, the magnetic core performs an AND function, makes a decision about whether to change the stored state, and stores state, all in one device. The MeRAM in Fig. 2B and the handling of the 9-core array as a single unit are examples of this principle.

I. While steps A-H merely quantify the limiting dissipation for a design, this quantity could be used as an objective function for design optimization. Logic design includes choices on how to encode information on wires and states. It also includes choosing amongst multiple gate-level implementations of a given function. In traditional logic design, these choices should all lead to correct designs that nonetheless vary in terms of speed, complexity, and energy consumption. However, the designs also differ in terms of minimum energy. If the designer is interested in the ultimate potential of a computing technology, the limiting dissipation computed in the steps above could guide a search for the design choices that yield minimum energy.

## VII. Conclusions

In this paper, we have described a path to reduced energy consumption in computers over the long term. Moore's Law and the principles of minimum energy for logic were properly stated in the 1960s, yet they are often interpreted specifically in context of CMOS microprocessors and generic Boolean logic gates. Within this narrow context, the theoretical efficiency limits are just $10^2$-$10^4$ beyond current technologies, which is not enough headroom to continue the long-term energy efficiency scaling that is part of Moore's Law.

We updated the example in Landauer's 1961 paper from an AND gate to a more modern synapse-like device and found a substantially lower theoretical bound on dissipation. A key difference is that our modern example exploits nonuniform input probabilities. The new theoretical bound may justify the perception that Moore's Law (defined for energy efficiency) can be extended further into the future than expected.

These ideas suggest research directions. One is the continued lowering of parasitic energy losses, $\underline{E}_r$ above. Another is a search for nanodevices that perform higher-level computations directly. These nanodevices would have lower energy dissipation than equivalent implementations using discrete gates, particularly if optimized for input statistics. There will be a need for many such nanodevices.

## References


[1] G. Moore, "Cramming more components onto integrated circuits," *Electronics* 38 (8), April 1965.

[2] D. J. Frank, "Power-constrained CMOS scaling limits," *IBM Journal of Research and Development* 46.2.3 (2002): 235-244.

[3] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. of Research and Development* 5.3 (1961): 183-191.

[4] C. Bennett, "Logical reversibility of computation," *IBM Journal of Research and Development* 17.6 (1973): 525-532.

[5] M. P. Frank, "Reversibility for efficient computing," Ph. D. thesis, Massachusetts Institute of Technology, 1999.

[6] R. W. Keyes, and R. Landauer, "Minimal energy dissipation in logic," *IBM Journal of Research and Development* 14.2 (1970): 152.

[7] M. P. Frank, "Towards a more general model of reversible logic hardware," *Superconducting Electronics Approaching the Landauer Limit and Reversibility* workshop, Annapolis, MD, Mar. 2012.

[8] N. G. Anderson, "On the physical implementation of logical transformations: generalized l-machines," *Theoretical Computer Science* **411**, 4179-4199 (2010).

[9] N. G. Anderson, "Irreversible information loss: fundamental notions and entropy costs," *International Journal of Modern Physics: Conference Series*, Vol. 33, 1460354 (2014).

[10] N. G. Anderson, "Conditional operations and Landauer's principle: aware erasure vs. oblivious erasure," in preparation.

[11] N. Ganesh, and N. G. Anderson. "Irreversibility and dissipation in finite-state automata," *Physics Letters A* 377.45 (2013): 3266-3271.

[12] J. Hu, et al. "High-density magnetoresistive random access memory operating at ultralow voltage at room temperature," *Nature communications* 2 (2011): 553

[13] V. K. Semenov, G. V. Danilov, and D. V. Averin, "Negative-inductance SQUID as the basic element of reversible Josephson-junction circuits," *Applied Superconductivity, IEEE Transactions on* 13.2 (2003): 938-943.

[14] J. Parrondo, J. Horowitz, and T. Sagawa, "Thermodynamics of information," *Nature physics* 11 (Feb. 2015): 131-139.

[15] P. Zulkowski and M. DeWeese, "Optimal finite-time erasure of a classical bit," *Physical Review E* 89.5 (2014): 052140.