Fundamental limits in dissipative processes during computation

Davide Chiucchiú¹, Cristina Diamantini², Miquel López-Suárez⁴, Igor Neri² and Luca Gammaitoni²

¹ Okinawa Institute of science and technology graduate university (OIST), Biological Complexity Unit. - 1919-1 Tancha, Onna, Kunigami District, Okinawa 904-0412, Japan.
² NiPS Laboratory, Dip. di Fisica e Geologia, University of Perugia - I 06100 Perugia, Italy.
³ INFN Perugia, via A. Pascoli, I-06100 Perugia, Italy.
⁴ ICMAB-CSIC, Barcelona, Spain.

* Correspondence: luca.gammaitoni@nipslab.org

Abstract: An increasing amount of electric energy is consumed by computers as they progress in function and capabilities. All of it is dissipated in heat during the computing and communicating operations and we reached the point that further developments are hindered by the unbearable amount of heat produced. In this paper we briefly review the fundamental limits in energy dissipation, as imposed by the laws of physics, with specific reference to computing and memory storage activities.

Keywords: dissipation; computing; fluctuations; heat; energy

1. Introduction

In the last fifty years the Information and Communication Technology (ICT) sector has experienced a huge growth, mainly fostered by the ability of the underlying semiconductor industry to repeatedly scale down the size of the CMOS-FET switches, the building block of present computing devices, and to increase computing capability density up to a point where billions of switches have been assembled in a square centimetre. Further progresses in this direction are thwarted by the amount of energy dissipated during switch operations: "the resulting power density for these switches at maximum packing density would be on the order of 1 MW/cm² - orders of magnitude higher than the practical air-cooling limit."[1]

There is little doubt that managing efficiently the use of energy, i.e. drastically reducing heat production, is a key aspect to consider in computing systems, especially for applications in smart sensors and Internet of Things devices, where the small dimension and the mobility require innovative solutions[2].

The energetic issues of future computers requires a clear understanding of their functioning in terms of efficiency, i.e. the amount of information processed per second, versus the amount of energy dissipated[3]. The search for an ultimate efficiency in computing is somehow similar to the search for the maximum efficiency in the functioning of heat engines that gave birth to the thermodynamics in the eighteen and nineteen centuries. However, at difference with the work done on steam engines aimed at reaching the ultimate limit set by the Carnot theorem, here a general agreement in the scientific community on the ultimate physical limits in energy dissipation during computation is still missing[4]. The controversy is mainly associated with the role to be assigned to the notion of information, as introduced by C. Shannon[5] in the early forties of the last century, and applied by Ralph Landauer[6] in the framework of what is nowadays called the Thermodynamics of computation. However, quite surprisingly, the notion of information, undoubtedly very useful in dealing with the mathematical
aspects of computing tasks, is not necessarily required when one is interested at the mere functioning of the machinery of the computing itself. As we will show in the following of this paper, we will carry on our analysis of the energetic efficiency just focusing on the functioning of the very basic physical elements of the digital computer: binary switches. In this perspective the physical description of the two main tasks of the digital computing process, i.e. logic-arithmetic operations and memory storage, can be performed without any recourse to the notion of information.

2. Binary switches

Automatic digital computing is performed by manipulating binary logic states encoded in physical devices. The most relevant devices employed in such a task are logic gates and memories. In binary logic, where the logic states are just two, often identified with logic state 0 and logic state 1, the state manipulation is performed according to the Boolean logic and arithmetic operations are realised by the repeated applications of basic logic operators called logic gates. All the logic and arithmetic operations, thus all digital computing, can be performed by assembling together sets of universal logic gates, like the NAND gate.

The NAND gate, whose logic output is 1 when and only when its two logic inputs are 0, and 0 otherwise, can be realised by employing two subsequent binary switches, as illustrated in Figure 1. In the role of binary switch it can be employed any device capable of assuming two different physical states (often denominated open and close) as a result of the application of an external force.

![Figure 1. Universal logic gate NAND. (Left) The logic output OUT is here associated with an electric voltage value across a simple circuit: high-V corresponds to the logic state 1 and low-V corresponds to the logic state 0. Binary switches are represented here as mechanical switches that can assume the logic state “0” (physical state open) or the logic state “1” (physical state close). When both the switches are in the close position the circuit conduces and the voltage $V_{OUT}$ position assumes the value low-V. (Right) Logic gate NAND implemented using transistors as binary switches. In this case both the input logic state and the output logic state are physically encoded into electrical voltages.](image)

In order to perform logic and arithmetic operations we need to change the state of a binary switch. In general this can be realised by the application of a generalised force $F$ that, by acting from outside on the switch, induces state changes, i.e. a switch event. We can call $F_{01}$ ($F_{10}$) the generalised force acting to bring the switch from 0 to 1 (1 to 0). Following this approach we can easily identify two different classes of devices that can be usefully employed as binary switches. We will call them combinational and sequential switches.

In a combinational switch we observe the following operating behaviour: when no external force is present, under equilibrium conditions, the switch is fund in the logic state 0. When an external force $F_{01}$ is applied, it switches to the logic state 1 and remains in that state as long as the force is applied. Once the force is removed it goes back to the logic state 0. A common example of combinational switch is the electro-mechanical relay, a switch used in many circuits involving lighting. Here when a magnetic induction force is applied, the relay changes its state from open to close and goes back to the
initial state (open) when the force is removed. Another important example of a combinational switch is the transistor, widely employed in modern computers where the input force is represented by an externally applied voltage that makes the transistor switch from the high-impedance (non-conducting) to low-impedance (conducting) condition.

On the other hand, in a sequential switch we observe a different behaviour: if the switch is in the logic state 0, it can be changed into the state 1 by applying an external force $F_{01}$, as before. However, in this case, when the force is removed, the switch remains in the logic state 1. This is true also for the switch event from 1 to 0 where the force $F_{10}$ has to be applied. We can say that, at difference with the combinational switch, the sequential switch remembers its state after the removal of the force. For such reason they are good candidates for realising storage devices and are widely employed in computer memories.

2.1. Energy dissipation in charge-based switch devices

As we have mentioned earlier, present computers are built using CMOS (Complementary Metal-Oxide Semiconductor) transistors, employed both as combinational and sequential switches. For these devices we face a number of dissipative effects that are due to their peculiar functioning and/or to the technology employed (semiconductor based, electric charge devices). For them, the most relevant source of dissipation during the switching[3] is associated with the charging and discharging of the electrical capacitances that are used to set the required voltage for the functioning of the transistor. If $C$ is the capacitance involved, the amount of energy dissipated per switch event is twice the energy stored in the capacitor and amounts to:

$$E_C = CV^2$$

where $V$ is the voltage across the capacitor plates. If the frequency of switching is $f$ we have a resulting dissipated power equal to

$$P_C = \alpha CV^2 f$$

where $\alpha$ is a coefficient that ranges from 0 (extremely smooth switching) to 1 (square wave switching). This quantity is usually refereed as dynamic dissipation.

An additional source of dissipation arises from the so-called static leakage due to the presence of subthreshold leakage phenomenon, i.e. small electric currents that occur between the source and drain of a transistor when it is in the subthreshold region and no conduction is expected. With the progressive reduction of the voltage $V$, operated with the aim of decreasing the dynamic dissipation due to the switching, the subthreshold leakage has increased its importance and with voltages as low as 0.2$V$, leakage can exceed 50% of total power consumption[7]. In absolute values the amount of energy dissipated during a single switch event has been continuously reduced over the last forty years and has presently reached the value of approximately $10^{-17}$J[3].

From the present description of the dissipation sources in charge-based switch devices, it appears clear that such phenomena are necessarily associated with the physical nature of the devices employed as binary switches. Perhaps, it should be expected that, once we change the physical device, for example substituting the electronic transistor with a nano mechanical cantilever, the dissipative mechanisms change as well. Thus, in order to inspect the fundamental limits in energy dissipation for computing devices, we should proceed with identifying a physical model of the binary switch and look for some dissipative mechanism that does not depend on the physical realisation of the switch itself and on the functioning principles being mechanical, optical, electro-mechanical or purely electronic.

2.2. The physics of binary switches

Let’s consider the general description of a binary switch, as previously introduced. This can be represented in terms of a one degree-of-freedom dynamic system, subjected to a confining potential
energy and external forces. If we assume $x(t)$ as the relevant variable, we can define two identifiable states, e.g. $x < x_{TH}$ (logic state 0), $x > x_{TH}$ (logic state 1), where $x_{TH}$ is an arbitrary threshold value. The variable $x(t)$ can represent an electric voltage (as in a transistor) or a position (as in a mechanical cantilever) or a value of the magnetisation (as in magnetic memories), just to mention few relevant examples. Here we are interested in those features of the variable $x(t)$ that are general enough to be common to all the cited cases. The time evolution of $x(t)$ can be described in terms of a second order differential equation, Langevin type:

$$\ddot{x} = -\frac{dU(x)}{dx} - \gamma \dot{x} + F(t) + \sigma \xi(t)$$

where $U(x)$ is the confining potential, $F(t)$ is the generalised force that produces the switch event, $\gamma$ is the dissipative constant, assumed a viscous damping-like friction, and $\xi(t)$ is a stochastic function with gaussian distribution, zero mean and unitary standard deviation. The presence of a non negligible stochastic force is required due to the fact that present binary switches have reached small physical dimension at the point that the presence of fluctuations cannot be neglected. At thermal equilibrium and in the presence of thermal noise as a dominant noise source, the spectral properties of $\xi(t)$ are set by the corresponding fluctuation-dissipation relation.

The dynamics represented in (3) can accommodate both the combinational and the sequential switches, depending on the specific choice of the potential function $U(x)$. Without lack of generality, here we will consider

$$U(x) = a \frac{1}{2} x^2$$

for the combinational switch and

$$U(x) = -a \frac{1}{2} x^2 + b \frac{1}{4} x^4$$

for the sequential switch. With $a$ and $b$ properly chosen constant parameters.

Due to the presence of the stochastic force, the system dynamics can be conveniently represented through the time evolution of the corresponding probability density $p(x,t)$ that can be statistically

---

**Figure 2.** Potential functions $U(x)$ with the associated probability densities $p(x,t)$. (Left column) Potential function $U(x)$ for the combinational switch. Here the threshold $x_{TH} = 1.2$. Upper graph: $p(x,t)$ associated with the logic state 0. Here $p_0 \simeq 1$ and $p_1 \simeq 0$. Lower graph: $p(x,t)$ associated with the logic state 1. Here $p_0 \simeq 0$ and $p_1 \simeq 1$. (Right column) Potential function $U(x)$ for the sequential switch. Here the threshold $x_{TH} = 0$. Upper graph: $p(x,t)$ associated with the logic state 0. Here $p_0 \simeq 1$ and $p_1 \simeq 0$. Lower graph: $p(x,t)$ associated with the logic state 1. Here $p_0 \simeq 0$ and $p_1 \simeq 1$. 

---

The dynamics represented in (3) can accommodate both the combinational and the sequential switches, depending on the specific choice of the potential function $U(x)$. Without lack of generality, here we will consider

$$U(x) = a \frac{1}{2} x^2$$

for the combinational switch and

$$U(x) = -a \frac{1}{2} x^2 + b \frac{1}{4} x^4$$

for the sequential switch. With $a$ and $b$ properly chosen constant parameters.

Due to the presence of the stochastic force, the system dynamics can be conveniently represented through the time evolution of the corresponding probability density $p(x,t)$ that can be statistically
computed from (3) or obtained as the solution of the associated Fokker-Planck equation[8]. The two
states, 0 and 1, are realised with probability respectively \( p_0 \) and \( p_1 \) \((p_0 + p_1 = 1)\) given by:

\[
p_0 = \int_{-\infty}^{x_{TH}} p(x,t)dx, \quad p_1 = \int_{x_{TH}}^{+\infty} p(x,t)dx\)

with \( x_{TH} \) a properly chosen threshold value. In Figure 2 we illustrate the two potentials \( U(x) \)
with the associated probability densities \( p(x,t) \) corresponding to the logic states 0 (left) and 1 (right).

We note that due to the presence of the fluctuating force, the stochastic process \( x(t) \) will oscillate
around the minima of the potential \( U(x) \). For the sequential potential in (5) this implies that \( x(t) \)
performs occasional random crossings between the two wells and, at equilibrium, with a symmetric
potential and zero average fluctuating force, it requires that \( p_0 = p_1 \).

3. Fundamental energy limits in binary switches

We are now interested in understanding what is the minimum energy required for operating
binary switches.

In order to fix our ideas, let’s consider the switch dynamics defined by (3). For what we have
seen so far, the logic switch event consists in the transformation of the probability density, such that
for the switch 0 to 1, the initial \( p_0 \approx 1 \) and \( p_1 \approx 0 \) become \( p_0 \approx 0 \) and \( p_1 \approx 1 \) while, vice versa for
the switch 1 to 0, the initial \( p_0 \approx 0 \) and \( p_1 \approx 1 \) become \( p_0 \approx 1 \) and \( p_1 \approx 0 \). It is clear that in order to
understand what is the minimum energy involved we have to take into account all the possible energy
processes associated with such a transformation. Thus, other than taking into account the potential
energy change and the role of friction, we have also to consider the change in the system entropy. Here,
the system entropy associated with the stochastic process \( x(t) \) can be computed according to Gibbs as:

\[
S = -k_B \int_{-\infty}^{+\infty} p(x,t) \ln p(x,t)
\]

with \( k_B \) being the Boltzmann constant. In order to understand what is the minimum energy
involved, we will consider the two switch families separately.

3.1. Combinational switches and logic gates

As we have anticipated above, combinational switches are the building blocks of logic gates that
are realised by assembling together one or more combinational switches. Specifically, we need two
switches to make a NAND gate that being a universal gate can be used in all the logic and arithmetic
operations. Accordingly the minimum energy required to operate a logic gate is identified once we
define what is the minimum energy required to operate the combinational switch.

We note that, logic states 0 and 1 have not the same energy due to the shape of the potential in (4).
If we indicate with \( \Delta E = E_1 - E_0 \) the potential energy difference, it is clear that in order to perform the
switch, the deterministic force \( F(x,t) \) has to perform a work at least equal to \( \Delta E \), against the potential
energy. What about the other forces? The stochastic force \( \xi(t) \) has zero mean thus, on average, the
work performed is zero. Finally the dissipative force \( \gamma \ddot{x} \) perform a dissipative work that is proportional
to the speed of the switch. In the adiabatic limit, i.e. when the switch event is produced with \( \dot{x} \to 0 \)
the dissipated energy tends to zero. Due to the presence of the fluctuations, we have to consider the
existence of a thermal bath that exchanges a certain amount of heat equal to \( T \Delta S \). However, due to
the linearity of the potential, it is always possible to choose a \( F(x,t) \) such that the system entropy
at the beginning and at the end of the switch is the same. Thus we have \( \Delta S = 0 \) and no minimum
net contribution from heat is required. Based on this analysis, it is clear that the work done by the
deterministic force \( F(x,t) \) during the 0 towards 1 switch it is the opposite of the work done during the
1 towards 0 switch.

In summary, if we are capable of performing the switch event with arbitrarily small velocity,
being the amount of work done during the 0 to 1 switch equal to the 1 to 0 switch and, being these
events on average equally likely in a long computation task, we can conclude that it is possible, at
least in principle, to perform such a task by spending zero energy. In order to better illustrate this
point we performed an experiment whose results have been discussed in ref. [9], where a micro
electro-mechanical cantilever has been employed as a combinational switch. In Figure 3 we show
the amount of energy dissipated during a cycle of two subsequent switches 0 to 1 to 0, performed
at different velocities. It is apparent that increasing the protocol time, the produced heat decreases
following a power law, in good agreement with friction model used to account for the velocity
dependent dissipation[9].

![Figure 3. Average heat production during the cycled combinational switch operation as function of
protocol time $\tau_p$. Increasing the protocol time the produced heat decreases following a power law.
The average heat $Q$ is shown here in $k_B T$ units, where $k_B$ is the Boltzmann constant and $T$ is the room
temperature.]

3.2. Sequential switches and memory devices

If combinational switches are the building blocks of logic gates, sequential switches are the basic
components of digital memories.

In this case two different operations have to be considered in order to account for the energy
dissipation during their functioning. The first operation is called reset and takes place when the
sequential switch is at equilibrium. In this case, in order to write a memory bit we need to break
the equilibrium condition and to re-set the binary switch in one of the two logic states. The second
operation is the switch and it is carried out when the initial state is known and one needs to change it.

3.2.1. The reset operation

As we have previously mentioned, the equilibrium condition of $x(t)$ according to (3) (5), is
represented by a symmetric probability density $p(x, t)$ where $p_0 = p_1 = 0.5$ (see Figure 4, left). In
order to store a binary digit we need to set our memory in a given state, 0 or 1. This operation requires
a change in the probability density that becomes entirely confined within one well of the potential.
In Figure 4 we show the change in $p(x, t)$ for the reset to 0 operation. Here we have to deal with the
heat exchanged with the thermal bath because the system entropy at the beginning and at the end of
the switch is not the same. It is easy to show that the required change in entropy is $\Delta S = -k_B \log 2$
and thus a minimum net contribution of $Q = k_B T \log 2$ is required. Such a work is realised by the
deterministic force. Finally, the dissipative work done by the frictional force can be reduced to zero if a
proper adiabatic protocol is followed.

3.2.2. The switch operation

The switch event is pictorially represented in the right column of Figure 2, where $p(x, t)$ associated
with the logic state 0 is shown in the upper graph. Here $p_0 \approx 1$ and $p_1 \approx 0$. The switch event consists
in changing the $p(x, t)$ into that one represented in the lower graph, associated with the logic state 1
where $p_0 \approx 0$ and $p_1 \approx 1$. At difference with the previous case of the combinational switch, here, the
logistic states 0 and 1 have the same energy thanks to the potential symmetry in eq. (5) and no net work
is required by the deterministic force $F(x, t)$. Once again particular care has to be devoted in selecting
a switch protocol that keeps the system as close to equilibrium as possible (adiabatic transformation)
in order to minimise the action of the dissipative force $\gamma \dot{x}$. We observe that this requirement can be
difficult to satisfy. As a matter of fact there are two conflicting requirements: on one side we want to
perform a switch that is slow enough to dissipate little or no energy at all and, on the other hand, we
want to perform the switch in a time that is shorter than the relaxation time that brings the $p(x, t)$ to its
relaxed state with $p_0 = p_1 = 0.5$. In the paper by Gammaitoni et al.\cite{10} a possible protocol capable of
satisfying such a requirement is illustrated together with some experiment performed by using nano
magnets. In summary, also in this case, the minimum heat exchanged with the thermal bath can be
zero, by the moment that the change in entropy is zero.

3.2.3. Memory preservation

Fundamental for the functioning of the memory is that the potential barrier allows to statistically
confine $x(t)$ for a given time within one of the two wells (Figure 2, right), hence ensuring that one
given bit is effectively stored. As we have anticipated, this confined state represents a non-equilibrium
condition that evolves, within the system relaxation time $\tau_k$, to statistical equilibrium (fig. 5). This
process is described via the time evolution of the probability density function $p(x, t)$ as follows. Let us
assume we have a memory where the bit 1 is stored. The initial probability density $p(x, 0)$ shows a
sharp peak centred in the right well (fig. 5, leftmost panel). According to the dynamics of the system,
$p(x, t)$ will first relax inside the right well and then it will diffuse into the left well, thus developing
a second peak. At any given time $t$, the probability that the system encodes the wrong logic state
is represented by $p_0(t)$ that grows towards the equilibrium condition $p_0 = 0.5$ when the memory is
statistically lost (fig. 5, rightmost panel).

In order to avoid memory loss a periodic refresh procedure is performed on any binary switch.
This procedure consists in reading and then writing back the content of the memory, and it is performed
at intervals $t_R$\cite{11}. The refresh operation restores a non-equilibrium condition by shrinking the width
of each peak of $p(x, t)$ back to its original state and requires some energy to be dissipated. This is due
to the fact that during this transformation, external work is done by the deterministic force in order
Figure 6. Experimental results of produced heat for a single refresh operation as function of the protocol time $t_p$. By increasing $t_p$ the produced heat tends to the lower bound $Q = -T \Delta S$ here represented by the dashed line. The solid squares represent the heat from the experiment, while the solid line is the fit with the Zener dissipative model.

To reduce the entropy associated with the increased width of the distribution. Such a work can be estimated by measuring the entropy change. In addition, we have to consider the losses associated with the friction if the transformation speed is not kept small enough.

The entropy change can be computed [12] by assuming that the dynamics of $x(t)$ is confined within one well and it can be approximately described by the dynamics of an harmonic oscillator, characterised by a Gaussian probability density function. Such an approximation is valid if $t_R$ is much smaller that the global relaxation time $\tau_k$. In this perspective the change in entropy can be computed as:

$$\Delta S = k_B \ln \left( \frac{\sigma_i}{\sigma_f} \right). \quad (8)$$

where $\sigma_i$ is the target standard deviation of the Gaussian peak to be achieved with the refresh and $\sigma_f$ is the standard deviation of the Gaussian peak before the refresh. While $\sigma_i$ can be arbitrary chosen, $\sigma_f$ depends on $t_R$ [12]. Hence the minimum required energy to operate a single refresh is $Q = T \Delta S$.

In order to test the practical attainability of this result we performed an experiment[12] by employing a micro electro-mechanical cantilever. A tiny NdFeB magnet is attached to the cantilever tip and an external electromagnet is placed in front of the cantilever in order to change the potential stiffness by changing the distance between the two magnets. The probability density change associated with the refresh operation is obtained by changing the stiffness of the potential.

In Figure 6 we show the measured values of $Q$ required to perform a single refresh operation as a function of the protocol time $t_p$, for fixed $\sigma_i$ and $\sigma_f$. We can see that $Q$ approaches the minimum value given by eq.(8) when $t_p$ increases towards the quasi-static protocol condition. In a practical memory, such a condition is clearly attainable only if the required protocol time satisfies the given condition $t_p \ll t_R \ll \tau_k$.

Provided that we want to keep the memory, i.e. preserving the stored bit with a probability of memory failure smaller than a given $P_E$, for a general time $\tilde{t}$, we computed[12] the minimum energy dissipation required as:
\[ Q_m = -N T \Delta S = \frac{T}{T_R} k_B T \ln \left( \frac{\sqrt{\sigma_i^2 + e^{-\frac{t_R}{\tau_i}}} (\sigma_i^2 - \sigma_w^2)}{\sigma_i} \right) \]  

(9)

where \( N \) is the number of subsequent refresh operations performed, \( \sigma_w \) and \( \tau_w \) are respectively the equilibrium standard deviation and the relaxation time inside one well. As it is well apparent, such an amount of energy dissipated can be, in principle, reduced to zero, provided that the probability distribution is kept constantly close to the equilibrium distribution and/or if the refresh time \( t_R \) is arbitrarily small.

4. Conclusions

In conclusion, we have briefly reviewed the fundamental limits in energy dissipation, as imposed by the law of physics, when basic computing tasks are performed. We have shown that logic gate operations, made by operating sets of combinational switches, can be, at least in principle, performed without any spending of energy. Same conclusion can be applied when sequential switches are operated as one-bit memory devices. The only exception is represented by the reset operation, necessarily required when a memory device is written for the first time. We have also shown that, once a one-bit memory has been written, its content can be kept for a given (finite) time \( t \) without spending any energy, provided the refresh operation is performed often enough so that the system does not change significantly its entropy.

Author Contributions: All the authors have equally contributed.

Funding: This research was funded by the European Commission under the following projects (FPVII, G.A. no: 318287, LANDAUER; G.A. no: 611004, ICT-Energy; G.A. no: 732631, OPRECOMP) and ONRG grant N00014-11-1-0695.

Conflicts of Interest: The authors declare no conflict of interest.

References