Chapter 7

Minimum Energy of Computing, Fundamental Considerations

Victor Zhirnov, Ralph Cavin and Luca Gammaitoni

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57346

1. Introduction

Energy consumption during computation has become a matter of strategic importance for modern ICT and its impact on future society. In this chapter we review some of the basic principles governing energy consumption in ICT and discuss future perspectives toward more efficient computers.

In the last forty years the progress of the semiconductor industry has been driven by its ability to cost-effectively scale down the size of the CMOS-FET [1] switches, the building block of present computing devices, and this has provided continuing increases in computing capability. However, this has been accompanied by a continuing increase in energy consumption and heat generation up to a point where the power dissipated in heat during computation has become a serious limitation [2, 3]. The energetics issues of current and future computation raises a question of the ultimate *energy efficiency* of computation that is reminiscent of the Carnot limit for the efficiency for heat engines [4]. It should be noted that the entire discipline of thermodynamics emerged from the practical need to increase the efficiency of utility heat engines. Innovations in steam engines and internal combustion engines have been driven by the need to more closely approach the ideal limit of a Carnot engine. Today, approximately 200 years after the work of Carnot, the problem of understanding the efficiency of generalized energy generators remains, although today the object of interest not only includes large power plants but also small scale devices and systems for information processing. In fact, one can view information processor as a *computing engine* that transforms incoming energy flow into useful work and also produces some heat [5].

Interesting insights on the energy efficiency of binary elements were obtained in pioneering studies by John von Neumann and subsequently by Charles H. Bennet and Ralph Landauer in the last century. It has been shown that information processing is intimately related to energy

management ("information is physical"). Specifically, Bennet and Landauer have shown that there exists a minimum amount of energy required to perform any irreversible computation. The ultimate limit on the minimum energy per switching is set at $k_B T \ln 2$ (approximately 10^{-21} J at room temperature) [6, 7] called the Shannon-von Neumann-Landauer (SNL) limit [8]. As Landauer argued, this minimum amount cannot be reduced to zero if some information is discarded (erased) during the computation process. The reason is directly linked to thermodynamics: erasing information decreases the overall entropy of the system and this cannot be done without dissipating heat of at least $k_B T \ln 2$ Joule per bit erased [7, 8].

While the physical limits of individual binary elements (switches) have been explored to a significant depth (many questions remain open however), currently, there are no theoretical results available that characterize the maximum computational efficiency of a computing systems as a whole; for example, in the spirit of the bound on efficiency of heat engines obtained by Carnot. A full understanding of possible limits to computational performance similar to the Carnot efficiency limit for heat engines would be extremely important both from theoretical point of view and could guide the design of future extremely energy-efficient computational engines. As an example, the Nanoelectronics Research Initiative was launched in 2005 [3], funded by a US-based consortium of semiconductor companies, and federal and state governments, to address a grand challenge to understand the fundamental energy limits of the physics of both binary elements (logic and memory) and computing systems. Before exploring the basic principles that govern minimum energy dissipation in devices and computation, it is appropriate to briefly review the state of the art for present computers.

2. Energy dissipation in present computers — A field survey

The four main information processing functions in modern electronic ICT systems are: computation, communication, storage, and display, as shown in Fig. 1 [9]. In the U. S. alone

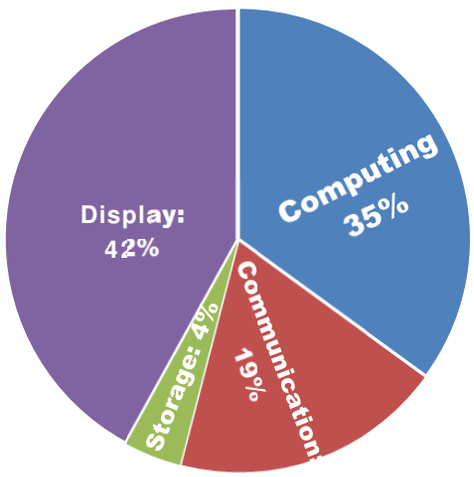


Figure 1. Energy consumption by four primary information processing functions in modern electronic ICT systems [9].

http://dx.doi.org/10.5772/57346

these constitutes about 290 TWh/year of electrical power, costing ~$30 Billion/year and the amount of electrical energy consumed by ICT is expected to continue to grow. It is instructive to review the sources of energy consumption in state-of-the-art ICT systems, since this may offer insights for possible directions to improve their energy efficiency.

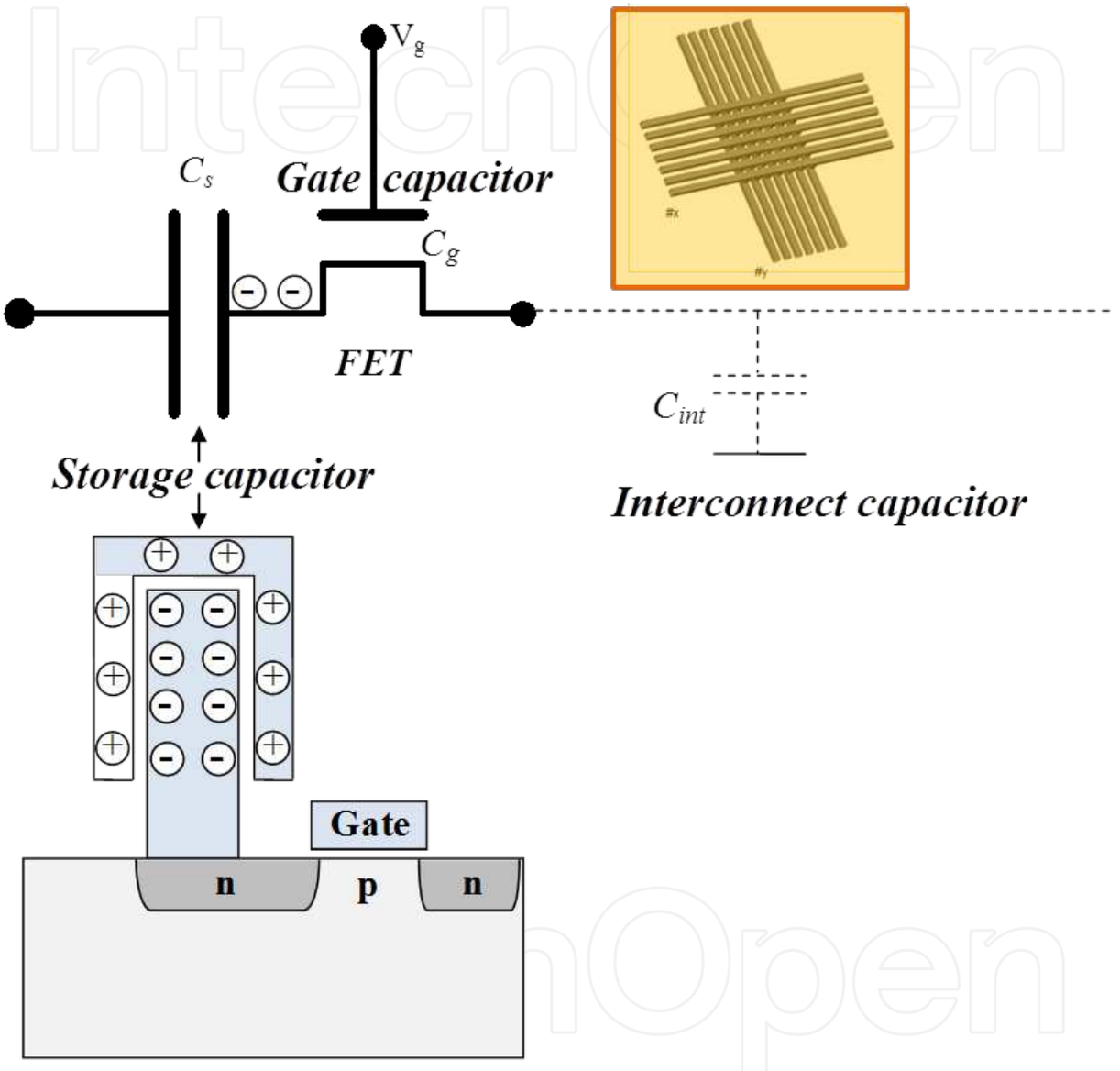


Figure 2. Device and circuit capacitance as a central concept of microelectronics.

2.1. Logic and memory devices

The main source of energy consumption in electronics is charging and discharging of electrical capacitances, which are present in all electronic devices. To illustrate this, an example of a dynamic random-access memory (DRAM) where several distinct capacitors are present, is shown in Fig.2 including a storage capacitor C_s, the gate capacitor C_g of the field effect transistor

(FET), and interconnect capacitor C_{int} formed by the wires used to connect individual memory cells in an X-Y array.

Energy dissipation by charging a capacitor is a central concept of microelectronics as operation of all electronic devices involves charging/discharging corresponding capacitors. When a capacitor is charged from a constant voltage power supply, energy is dissipated, i.e. converted into heat. Consider a typical model circuit consisting of a capacitor *C* in series with a resistor *R* (Fig. 3). Suppose a constant voltage of magnitude *V* is applied to the circuit at *t*=0 and electrical charge flows to the capacitor. The charging of the capacitor is characterized by a time-dependent voltage drops both on the resistor and the capacitor:

$$V_R(t) = V \exp\left(-\frac{t}{RC}\right) \tag{1}$$

$$V_C(t) = V\left(1 - \exp\left(-\frac{t}{RC}\right)\right) \tag{2}$$

The energy dissipated in the resistor *R* during charging is:

$$E_R = \int_0^\infty \frac{V_R^2(t)}{R}\,dt = \frac{V}{R}\int_0^\infty [\exp(-\frac{t}{RC})]^2\,dt = \frac{CV^2}{2} \tag{3}$$

Note that the energy dissipated in the resistor is independent of the resistance value *R*. As result of the charging process, the capacitor stores the energy $E_C = ½CV^2$, and thus the total energy required for constant charging voltage (the switching energy) is:

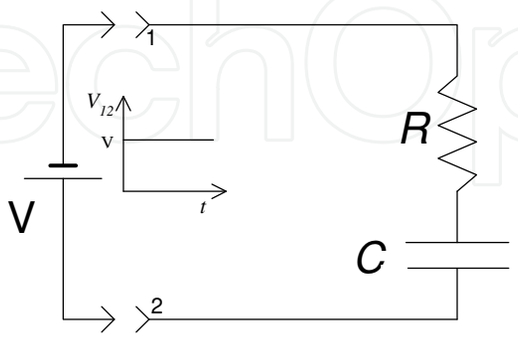


Figure 3. Generic *RC* circuit.

$$E_{sw} = CV^2 \tag{4}$$

http://dx.doi.org/10.5772/57346

Now capacitance is related to a linear dimension, L:

$$C \sim \varepsilon_0 L \tag{5}$$

If binary switching (i.e. capacitor charging and discharging) occurs with a frequency f, the operational power is:

$$P = \alpha E_{sw} f = \alpha C V^2 f \tag{6}$$

Where α is the activity factor, α=1, for a square-wave switching and α<0.5 in typical digital circuits. In a circuit with a large number of transistors, N_{tr}, the total switching energy is E_{sw} = N_{tr} E_{sw1}, where E_{sw1} is the switching energy of an individual transistor. Note that (4) and (6) refer to *dynamic* switching energy and power, directly related to the ON/OFF switching. There is also a parasitic leakage power component that will be discussed in the following.

According to (4), on the level of elementary operations (e.g. binary switching), the control space is limited by only two parameters – operating voltage and device and capacitance (devices and interconnects), and both these parameters have been considerably reduced during past 40 years as result of *scaling* – the continuing decrease of the device critical dimension, F, from tenths of micrometers to only a few nanometers. One immediate implication of scaling is a linear decrease of device capacitance (e.g. the FET gate capacitance, C_g) with F: C_g=k_1F. If voltage can also be linearly scaled with the device size with some coefficient of proportionality k_2, i.e. V=k_2F then the FET switching energy (CV^2) decreases as the cube of the device dimensions:

$$E_{sw1} = C_g V^2 = k_1 F \cdot \left(k_2 F\right)^2 = k_1 k_2 F^3 \tag{7}$$

Switching energies of individual transistors for several generations (1994-2011) of microprocessor units (MPU) are shown in Table 1 and Fig. 4. The data points for this 'bottom-up' approach were taken from several editions of the International Technology Roadmap for Semiconductors (ITRS) [10]. Note that the data points are approximated by a nearly cubic power function with strong correlation (the determination coefficient R^2 = 0.97), which is consistent with (7).

Now, it is instructive to compare the 'bottom-up' number with a 'top-down' average energy per transistor calculated from total power dissipation in practical microprocessors. From (6):

$$\left\langle E_{sw1} \right\rangle_{MPU} = \frac{P}{\alpha f N_{tr}} \tag{8}$$

The 'top-down' data points plot in Fig. 4 were obtained using (8), data from Table 1 and assuming the activity factor α=0.25.

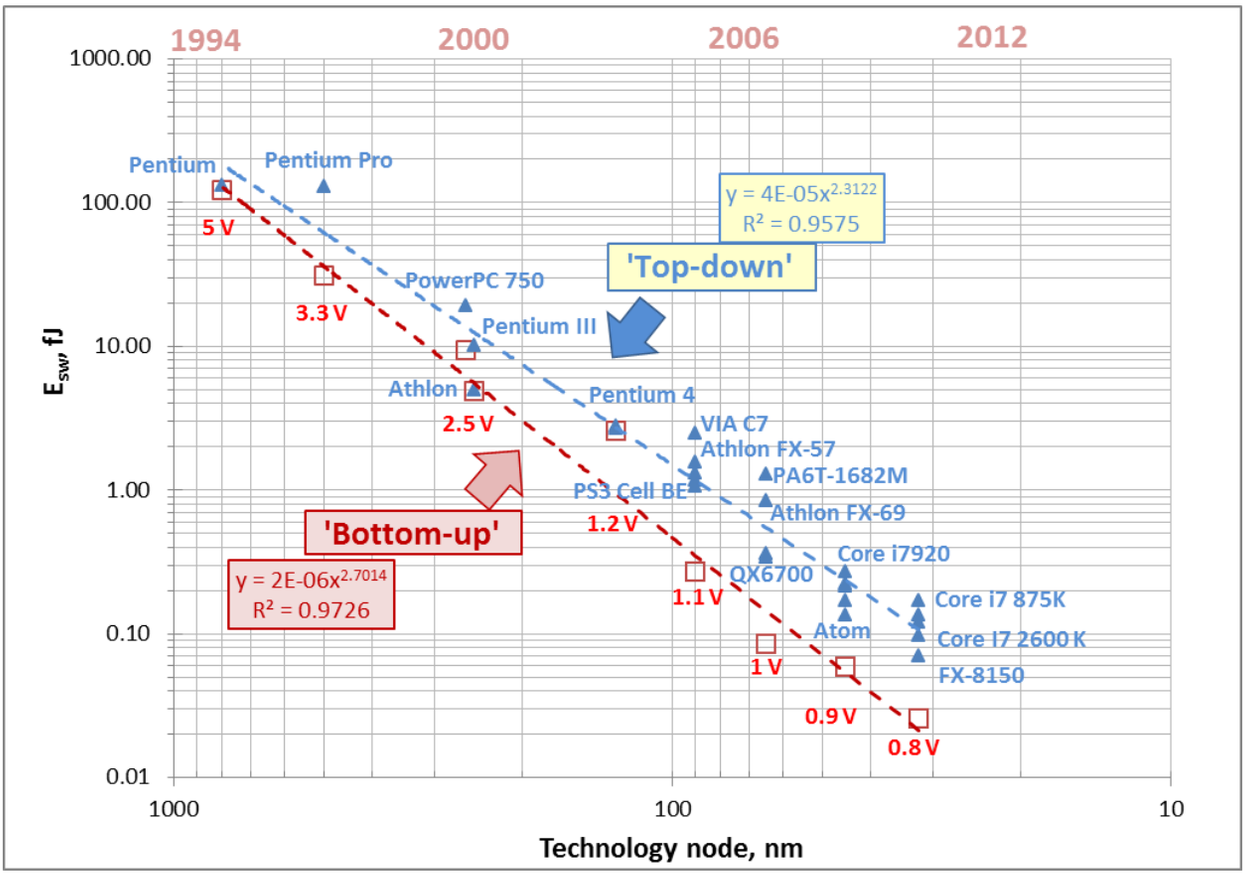


Figure 4. Switching energies of individual transistors in several generations (1994-2011) of microprocessors, calculated using 'bottom-up and 'top-down' approaches

Comparison of the 'top-down' and 'bottom-up' lines shows a clear divergence of the two as scaling continues. For a better visualization, a plot in Fig. 5 shows the ratio of the individual transistor dynamic switching energy ('bottom-up') to the average energy per transistor in MPU ('top down'): $\frac{E_{sw1}}{\langle E_{sw1} \rangle_{MPU}}$.

While for larger scale devices (e.g. ~ 1μm), the switching energy of transistor is a determining factor in the total chip energy consumption, for sub-100 nm technology nodes, the fraction of the transistor dynamic energy in the energy balance decreases. Indeed, transistor dynamic energy consumption constitutes ~20-30% of the total energy in modern microprocessors (22-65 nm node), and is expected to further decrease for 10nm devices and below. This trend is driven by increased dissipation in interconnects and off-state leakage losses.

As follows from the above, wires connecting binary switches, constitute a significant (and often dominant) portion of the energy consumption in ITC, and suggests a Carnot-type efficiency limit for computational engines.

http://dx.doi.org/10.5772/57346

Processor	Year	F, nm	IPS	f, MHz	N_{tr}	P, W	'bottom-up' E_{sw}, fJ	'top-down' E_{sw}, fJ
Intel Pentium	1994	800	1.88E+08	100	3.10E+06	10.1	122.72	130.32
Intel Pentium Pro	1996	500	5.41E+08	200	5.50E+06	35.0	31.32	127.27
PowerPC 750	1997	260	5.25E+08	233	6.35E+06	7.0	9.40	18.92
Intel Pentium III	1999	250	2.05E+09	600	9.50E+06	7.0	4.93	4.91
AMD Athlon	2000	250	3.56E+09	1200	2.20E+07	65.7	4.93	9.95
AMD Athlon XP 2500+	2003	130	7.53E+09	1830	5.43E+07	68.0	2.59	2.74
Pentium 4 Extreme Edition	2003	130	9.73E+09	3200	5.50E+07	115.0	2.59	2.61
VIA C7	2005	90	1.80E+09	1300	2.50E+07	20.0	0.27	2.46
AMD Athlon FX-57	2005	90	1.20E+10	2800	1.14E+08	104.0	0.27	1.30
AMD Athlon 64 3800+ X2 (Dual core)	2005	90	1.46E+10	2000	1.54E+08	89.0	0.27	1.16
Xbox360 IBM "Xenon" (Triple core)	2005	90	1.92E+10	3200	1.65E+08	203.0	0.27	1.54
PS3 Cell BE (PPE only)	2006	90	1.02E+10	3200	2.41E+08	200.0	0.27	1.04
AMD Athlon FX-60 (Dual core)	2006	65	1.89E+10	2600	2.33E+08	125.0	0.09	0.82
Intel Core 2 Extreme X6800 (Dual core)	2006	65	2.71E+10	2930	2.91E+08	75.0	0.09	0.35
Intel Core 2 Extreme QX6700 (Quad core)	2006	65	4.92E+10	2660	5.82E+08	130.0	0.09	0.34
P.A. Semi PA6T-1682M	2007	65	8.80E+09	2000	1.10E+07	7.0	0.09	1.27
Intel Core 2 Extreme QX9770	2008	45	5.95E+10	3200	8.00E+08	136.0	0.06	0.21
Intel Core i7 920	2008	45	8.23E+10	2660	7.31E+08	130.0	0.06	0.27
Intel Atom N270	2008	45	3.85E+09	1600	4.70E+07	2.5	0.06	0.13
AMD Phenom II X4 940	2009	45	4.28E+10	3000	7.58E+08	125.0	0.06	0.22
AMD Phenom II X6 1100T Thuban	2010	45	7.84E+10	3300	9.04E+08	125.0	0.06	0.17
Intel Core i7 Extreme Edition 980X	2010	32	1.48E+11	3330	1.17E+09	130.0	0.03	0.13
Intel Core i7 2600K	2011	32	1.28E+11	3400	1.16E+09	95.0	0.03	0.10
AMD E-350	2011	32	1.00E+10	1600	3.80E+08	18.0	0.03	0.12
Intel Core i7 875K	2011	32	9.21E+10	2930	7.74E+08	95.0	0.03	0.17
AMD FX-8150	2011	32	1.09E+11	3600	2.00E+09	125.0	0.03	0.07

Table 1. A 1994-2011 MPU summary

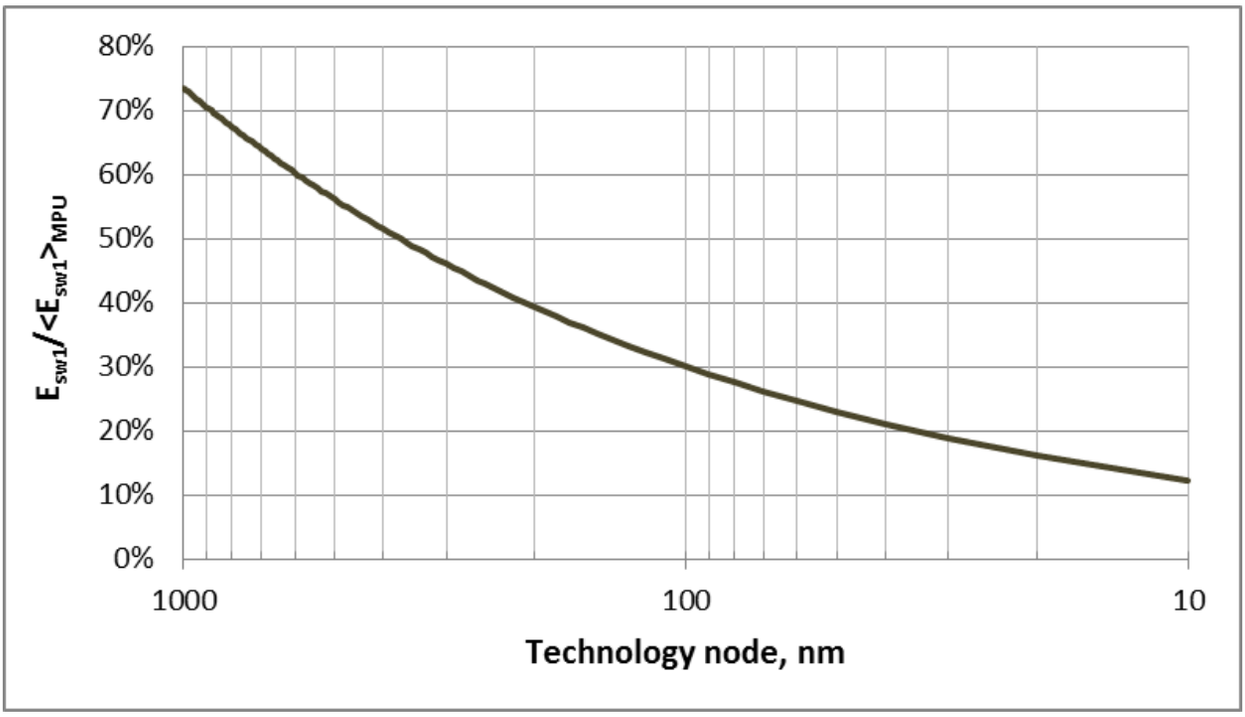


Figure 5. The ratio of the individual transistor dynamic switching energy ('bottom-up') to the average energy per transistor in MPU ('top down').

2.2. ICT Systems

One indicator of the ultimate performance of an information processor, realized as an interconnected system of binary switches, is the binary information throughput (BIT); that is the maximum number of on-chip binary transitions per unit time. BIT is the product of the transistor count N_{tr} with the clock frequency of the microprocessor f:

$$BIT = N_{tr} \cdot f \tag{9}$$

It is instructive to investigate its relation to the overall computational performance of microprocessors, which is often measured in (millions) of instructions per second (IPS) that can be executed against a standard set of benchmarks. As can be seen in Fig. 6, there is a strong correlation between system capability for IPS and the binary throughput, and to a good approximation:

$$IPS = k \times \left(BIT\right)^{r} \tag{10}$$

For a variety of microprocessor chips (a selection of 39 chips produced in 1971-2011 by 10 different companies, for details, see [11], k~0.1 and r~0.64 with a high degree of accuracy (the determination coefficient $R^2 = 0.98$). This strong correlation suggests a possible fundamental

http://dx.doi.org/10.5772/57346

law behind the empirical observation. It is also instrumental for speculations about future developments. According to (10), for a larger computational power, the binary throughput needs to be further increased, which in- turn requires an increase in the number of transistors and/or switching frequency. It is straightforward to show, however, that increasing BIT leads to increased power consumption, according to an equation:

$$P = BIT \times E_{bit} \tag{11}$$

Leading-edge high-performance chips already consume ~100 W of power (Figs. 6 and 7), and this makes their cooling an important issue.

A connection between binary information throughput and power consumption is very visible in memory blocks. Figure 6 shows a linear relation between read power and data rate for several high-speed DRAM systems, consistent with (11).

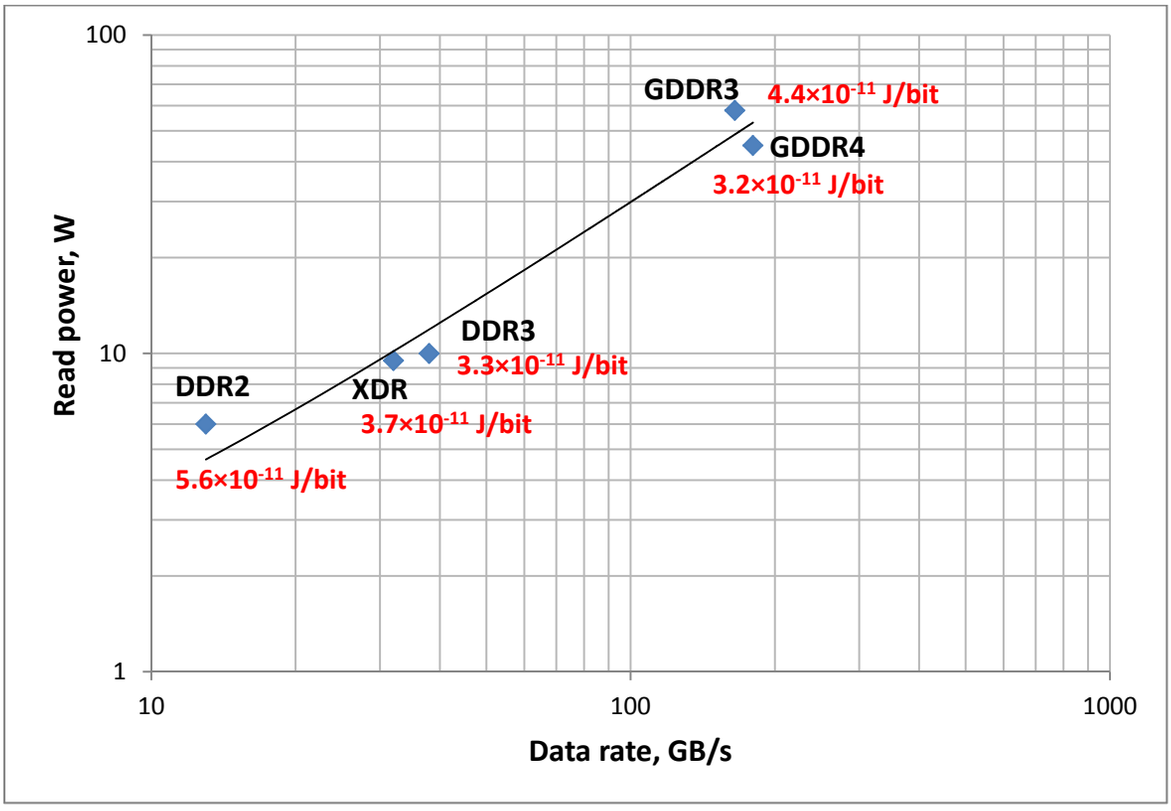


Figure 6. DRAM read power vs. data rate (adapted from [12])

The critical role of wires is also emphasized in memory circuits that are typically organized in arrays connected with long wires shown in an insert in Fig. 2. The wire capacitance (proportional to the wire length) effects system-level energy per bit operation (calculated using (11)

and indicated in red in Fig. 6). Note that the DRAM energy per bit is 10,000 times larger than the system-level energy per bit in microprocessors!

When searching for alternative information processing technologies and architectures, the human brain is often proposed as a different model for computation. In [13] an estimate of equivalent binary transitions was made from the analysis of the control function of brain: the equivalent number of binary transitions to support language, deliberate movements, information-controlled functions of the organs, hormone system etc., resulting in an 'effective' binary throughput of the brain ~ 10^{19} bit/s. An estimate of the number of equivalent instructions-per-second (IPS) was made in [14] from the analysis of brain image processing capability resulting in ~10^{14} IPS. It is clear that the brain is not on the microprocessor trajectory in Fig. 7, giving rise to the hope that there may exist alternate technologies and computing architectures offering higher performance (at much lower levels of energy consumption). On the other hand, achieving brain performance with existing ITC would require a massive increase binary throughput of the computing system, and this would also result in high power consumption. For example, the most recent and most impressive demonstration of an artificial intelligence

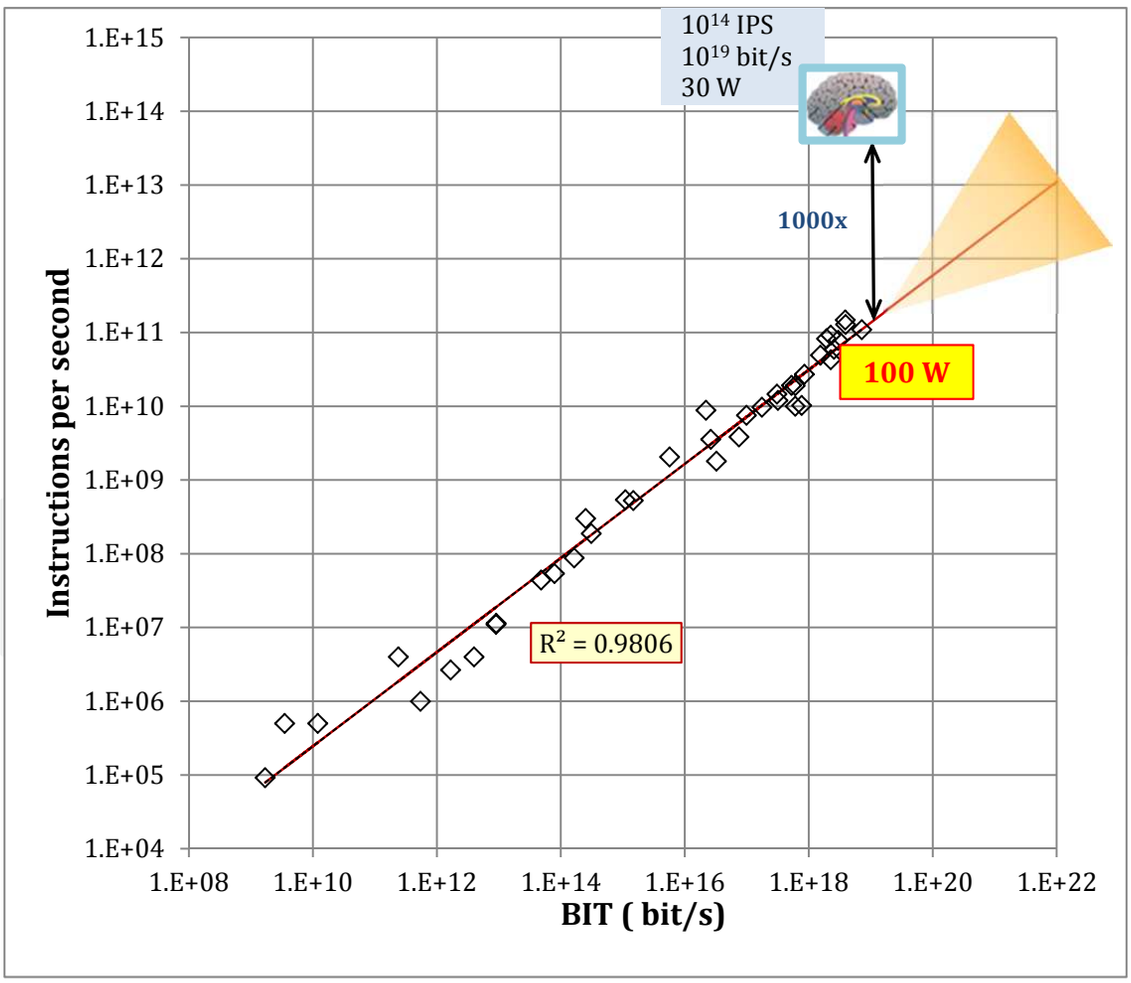


Figure 7. A trend for increasing computational performance through device scaling

http://dx.doi.org/10.5772/57346

computer system, the IBM Watson supercomputer capable of answering questions posed in natural language and the winner of the 2011 *Jeopardy!* quiz show, is built from ~3000 processor cores (POWER7) each consisting of 1.2B transistors and operating at 3.5GHz, thus approximate total binary throughput about 10^{22} bit/s. The machine consumes ~200kW of power [15]. The fact that the brain, a- biological information processor, operates at only ~30W suggests that there may exist alternate technologies and computing architectures offering higher performance (at much lower levels of energy consumption).

3. Fundamental limits in energy dissipation of computing

As we have seen in the previous paragraph, energy dissipation in present computers is an important issue. To reach a better understanding of the basic mechanisms behind energy dissipation in computing devices we propose to approach the energy dissipation issue from a very general and abstract perspective. We start this generalization by considering an ICT device as a black-box machine [16] that performs the activity of processing information by transforming some energy. For the moment we ignore any internal details of the functioning of this black-box. Under this perspective an ICT device can be considered a special thermal machine (see fig. 8).

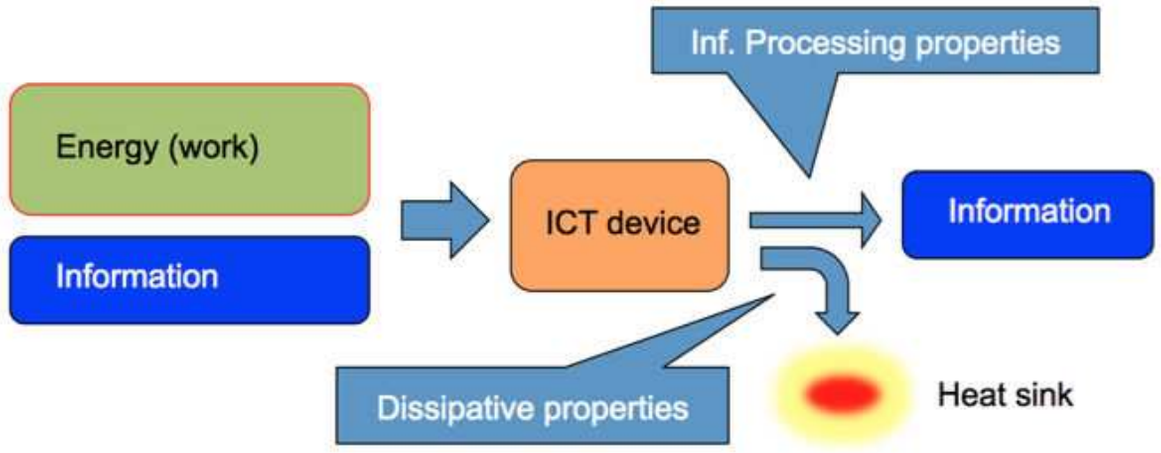


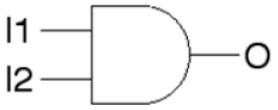
Figure 8. An ICT device is a machine that inputs information and energy (in the form of work), processes both and outputs information and energy (in the form of heat).

A traditional thermal machine is a device that processes energy. More precisely it transforms energy in the form of heat into work for industrial applications. An ICT device is a slightly more complex machine because it processes energy *and* information at the same time. More specifically it inputs a certain amount of information and some energy in the form of work and outputs a reduced amount of information and the same quantity of energy, although in the form of heat (see Fig. 8). In order to define the dissipative processes that take place during its functioning we need to consider both energy and information transformation processes. We

have already addressed the energy transformation processes in Chapter 2. Here we focus our attention on the information transformation processes.

3.1. Logic gates

In modern computers the information is processed via networks of *logic gates* that perform all the mathematical operations through assemblies of basic Boolean functions. As an example the NAND gate (Fig. 9) that, due to its universal character, can be widely employed in connected networks to perform a other logic functions. In Fig.10 the combinational networks of NAND gates for performing basic Boolean functions NOT, AND, OR are shown.



I1	I2	O
0	0	0
0	1	0
1	0	0
1	1	1

Figure 9. Symbolic representation of a NAND logic gate and the corresponding truth table. I1 and I2 are the input bits and O is the output bit.

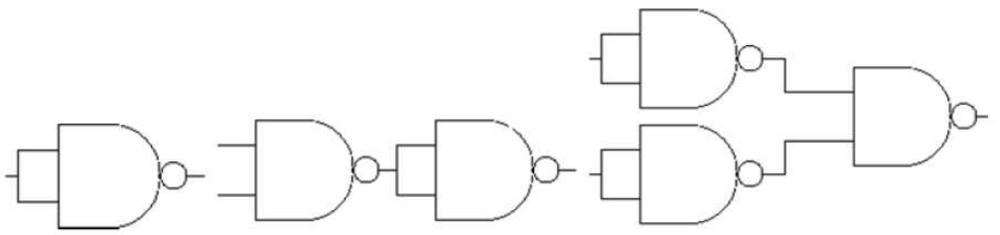


Figure 10. From left to right: NOT gate, AND gate, OR gate, all implemented by connecting together NAND gates.

According to the information preserving role of the logic operations we can distinguish the logic function in *logically reversible* logic gates and *logically irreversible* logic gates. For example the NOT function is implemented by a logically reversible logic gate because the value of the input bit I can be deduced by the value of the output bit O, as it is well evident by inspecting its truth table:

I	O
0	1
1	0

Figure 11. Truth table for the logic gate NOT.

http://dx.doi.org/10.5772/57346

On the other hand the NAND gate (see Fig. 9) is clearly logically irreversible because the knowledge of the output bit O does not allow to one to deduce the value of the input bits I1 and I2 in three cases out of four. According to the definition of *quantity of information* proposed by Shannon[17], an irreversible logic gate decreases the quantity of information at its output. As a simple example consider the NAND gate: there are two bits of information at the input (see truth table in fig. 9) and 1 bit of information at the output. Thus the information balance is negative and the logic gate is irreversible. On the other hand, in the case of the NOT gate there is 1 bit at the input and 1 bit at the output. The information balance is zero and the logic gate is reversible.

This section is entitled "Fundamental limits in energy dissipation of computing" but to date the focus has been on logic gates, i.e. mathematical operations. What has all this to do with energy? To answer this question we have to consider the fact that in a practical computer, the logic gate function is realized by some material system. The bit value is represented by some physical entity (signal) like electric current or voltage, light intensity, magnetic field,…etc. Such signal inputs to the logic gate device and go through a transformation to represent at the output the desired bit values. Modern logic gate devices are made by assembling more elementary units: i.e., transistors. A transistor is an electronic device that performs the role of a switch by letting or not-letting the electric current pass through. Examples of physical implementations of logic gates will be discussed below.

3.2. Landauer limit

In the following section it will be demonstrated that the minimum energy to operate a physical switch can be reduced to zero provided that the amount of information in the switch transformation is not decreased. This condition has been pointed out initially by John von Neumann in a lecture in 1949 [6] and subsequently focussed by R. Landauer [7] and C. H. Bennet[8].

The reasoning is the following: the switch is a macroscopic apparatus composed by many elementary parts (atoms) and thus can be considered a thermodynamic system approximately at equilibrium with the environment. As was shown above, this implies that its transformations are subjected to the laws of thermodynamics. We focus our attention on the single degree of freedom (*dof*) represented by the switch status. This is a dynamical system coupled to the thermal bath represented by all the remaining internal *dof*. A switch event is a change from an initial condition to a final condition. During this change the exchanges of energy and entropy need to be accounted for. If the switch is thermally isolated from the external environment, then there can be no transfer of heat. Suppose for a moment that a switching event can be performed without any work from outside (this point is addressed in the next paragraph), then the only balance that needs to be taken into account is the change in entropy. This change is measured by the change in the macroscopic configuration of the switch. If the change is from state *open* to state *close,* then there is only one initial configuration and one final configuration. There is therefore no net change in the number of configurations and thus no change in entropy according to Boltzmann (see Chapter 2). Let's suppose now that the switch is in an unknown state (it will be shown later that this is the natural condition for a physical switch left alone, after some time), this means that the switch can be in the *open* or in the *closed* state with equal

probability in the initial configuration. If now a change is applied to put the switch into a *close* (or *open*, same reasoning) condition, the number of configurations is changed from 2 to 1 and thus there is a change in entropy (Chapter 2) given by:

$$S_f - S_i = k_B(\ln 1 - \log 2) = -k_B \ln 2 \tag{12}$$

where K_B is the Boltzmann constant. The change in entropy is associated with a change in heat via the relation discussed in chapter 2:

$$TdS \geq dQ \tag{13}$$

The equal sign holds when there is no other dissipation associated with the change. Thus based on this reasoning every time that the number of input configurations is smaller than the output configurations, there is a reduction of entropy. Due to the second principle of thermodynamics, this process cannot occur spontaneously and energy expenditure is required. This energy has a lower bound in the amount just given above.

This result can be generalized to ICT devices composed of an arbitrary number of switches. The number of *input configurations* in a network of switches is associated with the *number of input bits* to a system of ICT devices and the number of *output configurations* is related to the number of *output bits*. Thus by computing the quantity of information change during the operation the minimum energy expenditure for the operation can be determined. For the simplest case (sometimes addressed as the *reset operation*) where a switch is set to a given value (*open* or *close*), the minimum energy amounts to:

$$k_B T \ln 2 \tag{14}$$

as anticipated at the beginning of this chapter.

The detailed physics of real switches is discussed in the following sections. The concentration of the following discussion is on energy dissipation and addresses the more fundamental question pertaining to the minimum energy dissipation in *any possible ideal switch*. In this regard a switch is an ideal device that can assume only two states: *open* and *close*.

In the following the focus will be on the physics of a switch with the aim of elucidating the general features associated with energy dissipation mechanisms that take place during the switch operation. In doing this, however, we will ignore those mechanisms that are associated with a specific technology (like the charging or discharging of a capacitor in the electronic realization of a switch) and try to discuss the mechanisms that are common to any possible realization of a physical switch. In order to reach this goal let's start with the definition of switch that we have introduced above: a (bistable) switch is a device that can assume two distinguishable states.

http://dx.doi.org/10.5772/57346

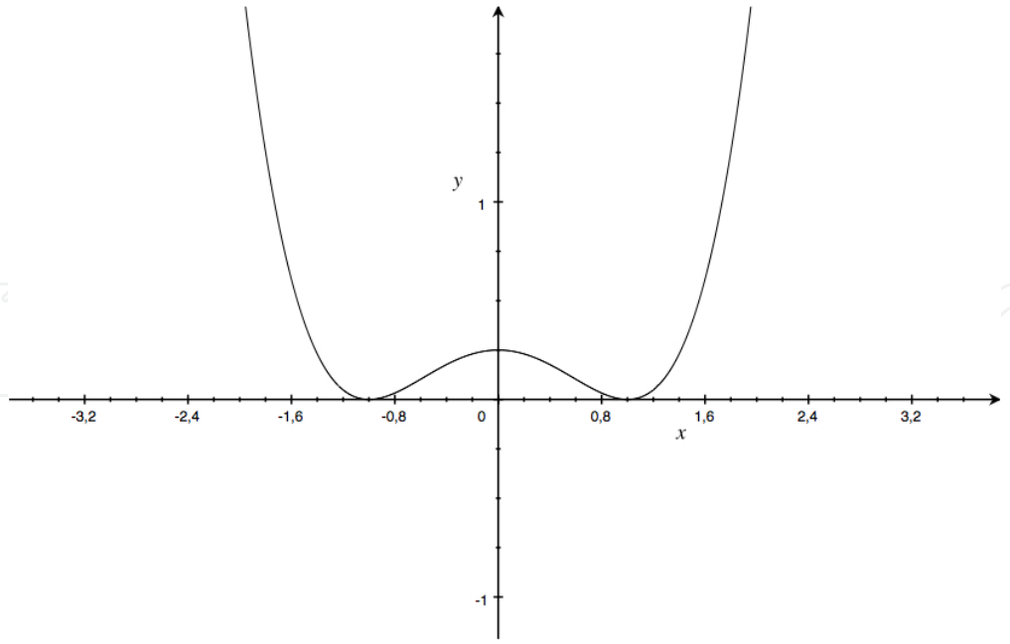


Figure 12. Bistable potential *U (x)*

3.3. The dynamics of a "simple switch"

In order to describe the physics of a switch we need to introduce a dynamical model capable of capturing the main features of a switch, regardless if it is realized with a purely mechanical, electro-mechanical or electronic technology. According to the reasoning originally developed by Landauer[7] we assume that the switch dynamics can be described by a single degree of freedom (*dof*) that is identified with x. Let's suppose that x is a continuous variable (e.g. the position of a cursor or the value of a magnetic field) that can assume two identifiable stable states: e.g. x<0 (logic state "0"), x>0 (logic state "1"). The two states, in order to be dynamically stable, are separated by some energy barrier that should be surpassed in order to perform the switch event. This situation can be mathematically described by a second order differential equation like:

$$m\ddot{x} = -\frac{d}{dx}U(x) - m\gamma\dot{x} + F \tag{15}$$

Where F is an external force that can be applied when we want to change state, γ is the frictional force that represent dissipative effects in the switch dynamics and

$$U(x) = -\frac{1}{2}x^2 + \frac{1}{4}x^4 + c \tag{16}$$

is the bistable potential shown in fig. 12. The additive constant, c, is an arbitrary constant that sets the zero level of the potential energy.

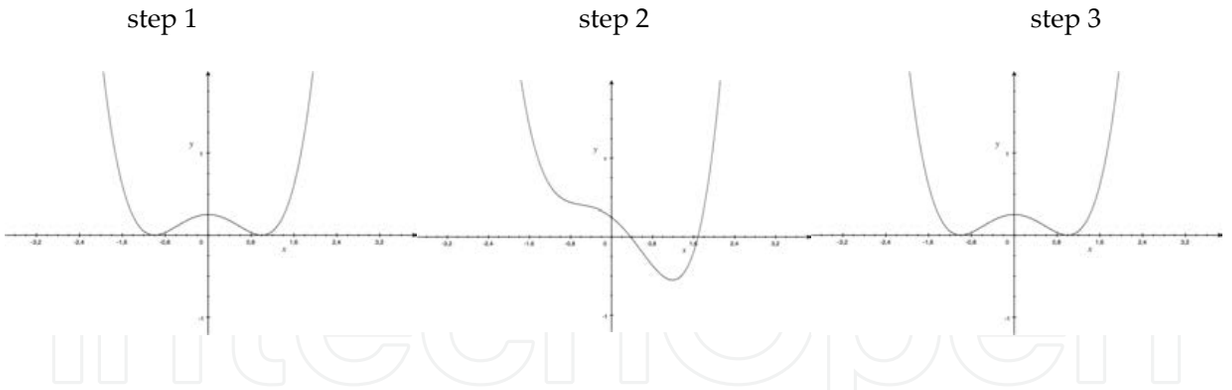


Figure 13. Potential $U(x) + F$. First procedure: From left to right, step 1,2,3. Step 1 and step 3, F=0; step 2, $F=-F_0$.

Suppose that at a certain time t_0, the system is x<0 (logic state 0) and $F = 0$. This is equivalent of picturing a material particle of mass m and position x, sitting at rest at the minimum of the left well in figure 15 and is an equilibrium condition for the switch.

According to this model if a switch event is to be produced it is necessary to apply an external force F capable of bringing the particle from the left well (at rest at the bottom) into the right well (at rest at the bottom). Clearly this can be done in more than one way.

As an example we start discussing what we call the **first procedure**: a three-step procedure based on the application of a large and constant force $F=-F_0$, with $F_0 >0$.

We start in step 1 (see fig. 13) with the particle on the left well and $F=0$. In step 2 we apply for a certain time $F=-F_0$ in order to change the potential shape into $U(x) - F_0 x$ (see fig. 13, step 2).

Clearly after some time the particle will move toward the right until it reaches the bottom of right well, where, after few oscillations, it settles due to the presence of the dissipative force. Then, step 3, the force F is removed and the system returns to the unperturbed potential of Fig. 12. In this way, a switch event can be produced.

What is the minimum work that the force F must perform to make the device switch from 0 to 1 (or equivalently from 1 to 0). The work is computed as:

$$L = \int_{x_1}^{x_2} F(x)dx \tag{17}$$

where x_1 and x_2 are the starting and ending position of our particle.

In the above example the work is readily computed by considering that the total force acting on the particle is $F_0 + dU/dx$ and has caused a displacement from x_1=-1 to x_2=1. The total work performed is easily computed to be $L_0 = 2 F_0$. Is this the minimum work? Clearly it is not.

In order to demonstrate that it is possible to switch with a less work, let's consider the following 5-step procedure (**second procedure,** see Fig. 14): in step 1 and step 5 let $F=0$; in step 2 lower the potential barrier by applying a proper force $F=-x$. In step 3 apply an additional small

http://dx.doi.org/10.5772/57346

constant force $-F_1$ that tilts the potential toward the left. Now $F=-x-F_1$. At this point the material particle slowly moves toward the right. When the particle reaches the far right limit proceed to step 4 and remove the $F=-x$ force. Finally in step 5, remove the additional force $F=F_1$ and restore the original bistable potential.

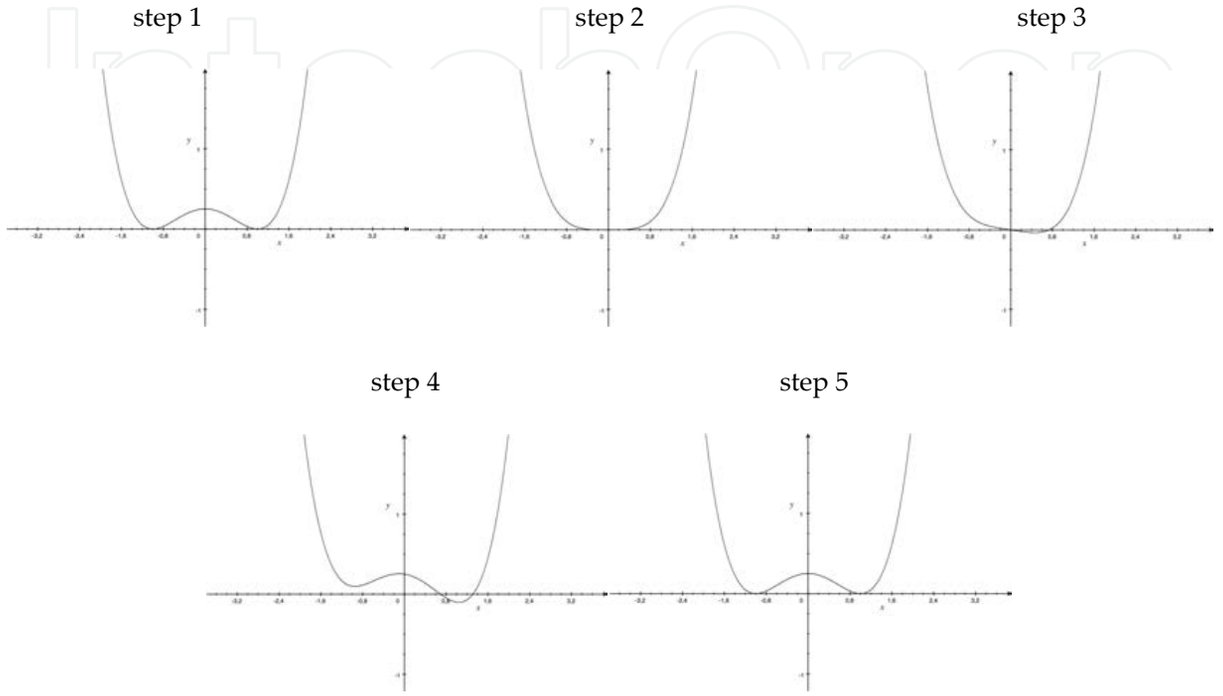


Figure 14. Potential $U(x) + F$. Second procedure: Step 1 and step 5, F=0; step 2 $F=-x$; step 3 $F=-x-F_1$; step 4 $F=-F_1$;

In order to compute the work performed on the particle observe that in step 1-2 and step 4-5 no work is performed because the applied force does not produce any displacement (or a negligible one). The only work performed happened to be during step 3 where it is readily computed as $L_1 = 2\,F_1$. Now, by the moment that $F_1 << F_0$, as anticipated, we have $L_1 << L_0$. Based on this reasoning it can be concluded that, provided an arbitrarily small constant force is applied during the tilt, the resulting work will be arbitrarily small. Thus it can be concluded that in principle it is possible to perform the switching event by spending zero energy provided two conditions are satisfied: 1) The total work performed on the system by the external force has to be zero. 2) The switch event must proceed with a speed arbitrarily small in order to have arbitrarily small losses due to friction.

3.4. The dynamics of a more realistic "simple switch"

This analysis, although correct, is quite naïve, indeed. The reason is that it has been assumed that the work performed, no matter how small, is completely dissipated by the frictional force. As we have discussed in chapter devoted to energy however, for an isolated system the existence of a dissipative force is the signature of the presence of a large number of degrees of freedom that somehow accommodates the dissipated energy associated with work done by

the force. In order to take into account a more realistic representation of the switch dynamics [18] assume that the single-*dof* switch is coupled to a thermal bath that is at temperature T. Although the switch is thermally isolated, exchanges of heat Q between the switch and the thermal bath are possible. Moreover, due to the coupling with the thermal bath a fluctuating force $\xi(t)$ appears. At thermal equilibrium the Fluctuation-Dissipation theorem (see Chapter on energy) links $\xi(t)$ and the dissipative force. According to this (more physical) description the switch dynamics can be described in terms of a Langevin equation, where the fluctuating force now appears:

$$m\ddot{x} = -\frac{d}{dx}U(x) - m\gamma\dot{x} + \xi(t) + F \tag{18}$$

The fluctuating force $\xi(t)$ is represented here by a zero average stochastic process that is defined in statistical terms. The equation of motion has now become a stochastic dynamical equation and its *solution* can be approached in statistical terms. One relevant quantity for describing the system dynamics is represented by the probability density function $P(x,t)$. Specifically $P(x,t)dx$ represents the probability for the observable x (the position of the particle) to be at time t within the interval between x and x+dx. Accordingly

$$p_0(t) = \int_{-\infty}^{0} P(x,\ t)dx \ \textit{and}\ \ p_1(t) = \int_{0}^{+\infty} P(x,\ t)dx \tag{19}$$

represent the probabilities for the switch to assume the logic states 0 and 1, respectively. As discussed before, it is now possible to address the problem of the work required to perform a switch event in this new thermodynamic framework.

In this case the definition of the switch event itself must be reconsidered. Previously the switch event was defined as the change from an equilibrium position (e.g. at rest at the bottom of the left well) to another equilibrium position (e.g. at rest at the bottom of the right well). In this new thermodynamic framework however the particle is never at rest: due to the presence of the fluctuating force the particle will be randomly oscillating around the potential minima, with occasional random crossings of the potential barrier between the two wells. Since the potential is symmetrical and the fluctuating force has zero average, the two states "0" and "1" have the same probability. This implies that the probability density distribution at equilibrium $P(x,t)=P(x)$ is stationary and symmetric, as represented in fig. 15.

Thus if the particle is placed at rest at the bottom of the left well, then after some time t_1 it starts to oscillate around the potential minima and after some longer time t_2 it will jump into the right well and eventually back into the left well and so on. The time t_1 and t_2 are random variables. Their mean values $\tau_1 = \langle t_1 \rangle$ and $\tau_2 = \langle t_2 \rangle$ (with $\tau_2 > \tau_1$) can be computed on the bases of the features of the potential $U(x)$ and the stochastic force $\xi(t)$. They are usually addressed as the *intra-well* relaxation time and the *inter-well* relaxation time and, roughly speaking they represent respectively the average time the system takes to establish equilibrium within one well (as it would temporarily ignore that the potential is wider than a single well) and the average time it takes to go to global equilibrium. Since τ_2 depends exponentially on the barrier

http://dx.doi.org/10.5772/57346

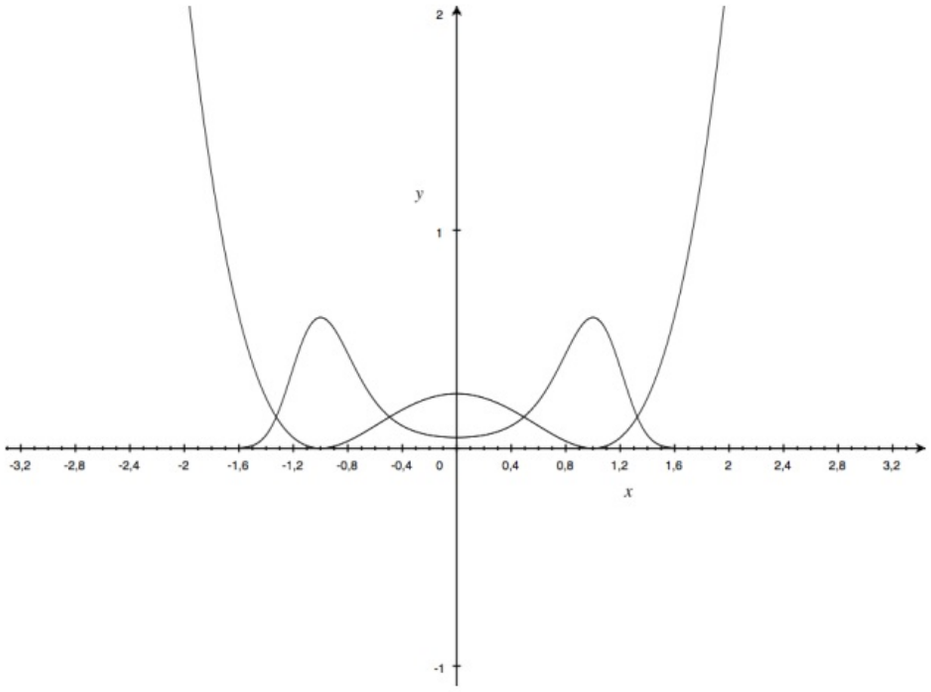


Figure 15. Bistable potential *U (x)* with superimposed the probability distribution P (x,t)=P (x) at equilibrium.

height between the two wells, in practical switches the barrier height is chosen to be large enough to guarantee that $\tau_2 >> \tau_1$.

Based on these considerations define the switch event as the transition from an initial condition toward a final condition, where the initial condition is defined as $<x> < 0$ and the final condition is defined as $<x> > 0$. With the initial condition characterized by:

$$p_0(t)=\int_{-\infty}^{0}P(x,\ t)dx \cong 1 \ and \ \ p_1(t)=\int_{0}^{+\infty}P(x,\ t)dx \cong 0 \tag{20}$$

and the final condition by:

$$p_0(t)=\int_{-\infty}^{0}P(x,\ t)dx \cong 0 \ and \ \ p_1(t)=\int_{0}^{+\infty}P(x,\ t)dx \cong 1 \tag{21}$$

Clearly the conditions are reversed for a switch event from 1 to 0.

In order to produce the switch event, we proceed as follows: set the initial position at any value $x < 0$ and wait a time t_a, with $\tau_1 << t_a << \tau_2$, then apply an external force F for an elapsed time t_b to produce a change in the $<x>$ value from $<x> < 0$ to $<x> > 0$. Then remove the force. In practice it will be necessary to wait a time t_a after the force removal in order to verify that the switch event has occurred, i.e. that $<x> > 0$. The total time spent has to satisfy the condition $2\ t_a + t_b << \tau_2$.

Now that a switch event has been defined in this new framework, we can return to the question: what is the minimum energy required to produce a switch event?

It is quite easy to see that in order to minimize the energy dissipation the role of the friction has to be negligible. This requires that the switch process must be performed very slowly. As already illustrated in the first procedure described previously (the constant force F_0 procedure) is not optimal. What about the second procedure? We can show that the second procedure, at difference with our previous analysis, although it does allow for a zero work transformation, it does not allow for zero energy expenditure. This is well apparent since in this new thermodynamic framework account must be taken for not only the energy changes due to the external work but also the heat Q passages and thus the role of the entropy S of the system. Based on the discussion in Section 11.3 and more generally in Chapter 2, we have seen that while a switching transformation can be carried-out without spending any energy, a transformation that evolve spontaneously (and thus increases the system entropy), if it is desired to perform a transformation that decreases the system entropy it is necessary to expend a minimum amount of energy $\Delta Q = T \Delta S$. In this case (particle in the double well) the system entropy can be computed according to Gibbs as:

$$S = - k_B \sum_i p_i \log p_i \tag{22}$$

Here the sum is limited to the two possible states in our switch and thus i=0,1. Thus if to perform a switch event without spending energy it is necessary to follow a procedure that does not require any entropy decrease during any of the steps. Let's analyse the steps in the second procedure. Note that between step 1 and step 2 the entropy of the system increases. This is due to the fact that the potential is changed by lowering the barrier. At this point the particle dynamics relaxes (in a very short time) to the new configuration and the entropy increases. This is apparent by the change in the probability distribution (see fig. 16) and can be demonstrated quantitatively by simply assuming that in step 1 we have $p_0=1$ and $p_1=0$, this gives $S_1 = -k_B \ln 1 = 0$. In step 2 $p_0=p_1=½$ (see fig. 19) and thus $S_2 = -k_B (½ \ln ½ + ½ \ln ½) = k_B \ln 2$. Thus $\Delta S = k_B \ln 2 > 0$. On the other hand, when there is a transition from step 2 to step 5 entropy is reduced from S_2 to $S_5=S_1=0$, thus $\Delta S = -k_B \ln 2 < 0$. According to the thermodynamics theses last steps cannot be performed without providing energy to the system and thus the minimum energy in this case is not zero.

Based on these considerations the conditions required to perform a switching event that expends zero energy can be formulated: 1) The total work performed on the system by the external force has to be zero. 2) The switch event has to proceed with a speed arbitrarily small in order to have arbitrarily small losses due to friction. 3) The system entropy must never decrease during the switch event.

In the following, as an example, a possible procedure (**third procedure**) that satisfies these three conditions is shown. In order to satisfy condition 1), apply a force that keeps the average position of the particle always close to the minimum of the potential well. In this case in fact the force is zero and the work will be zero as well. In order to satisfy condition 2) a change in

http://dx.doi.org/10.5772/57346

the applied force should be produced very slowly. Finally in order to satisfy condition 3), i.e., the probability density in state 0 and in state 1 is the same, apply a force that does not change the probability density along the path (constant entropy transformation). This can be done by applying a force that changes the potential as shown in fig. 17. Such a procedure clearly satisfies the three conditions that we enunciated above.

Finally, to conclude this section observe that any physical bistable switch, if left alone for a time that is of the order of τ_2 will eventually evolve into a situation similar to Fig. 15. In this case, when a switch event is required, the operation is completely similar to the reset operation addressed by Landauer in his original works and thus a minimum of k_B T ln 2 is necessarily required to operate the switch.

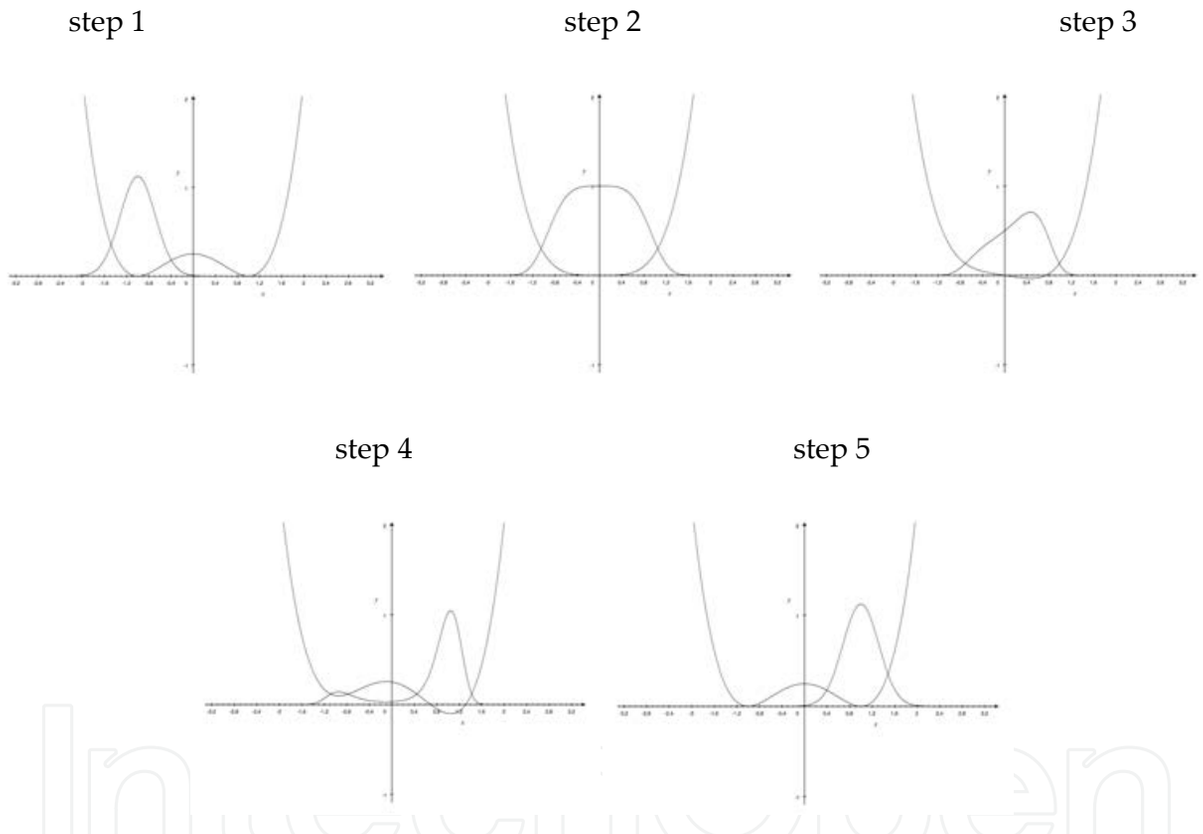


Figure 16. Potential *U (x)* + *F*. Equilibrium P (x) for the different cases. Second procedure: Step 1 and step 5, F=0; step 2 *F*=-x; step 3 F=-*x*-F_1; step 4 *F*=-F_1;

4. Charge based switching devices

Simplest charge based switch (Fig. 18) is a electromechanical device consisting of two metal electrodes, that, depending on the switch's state, are either separated by an air gap (OFF or *open* – non-conducting state) contacts, or touching each other (ON or *closed* – conducting state). The separation between electrodes is changed by applying external mechanical force (e.g.

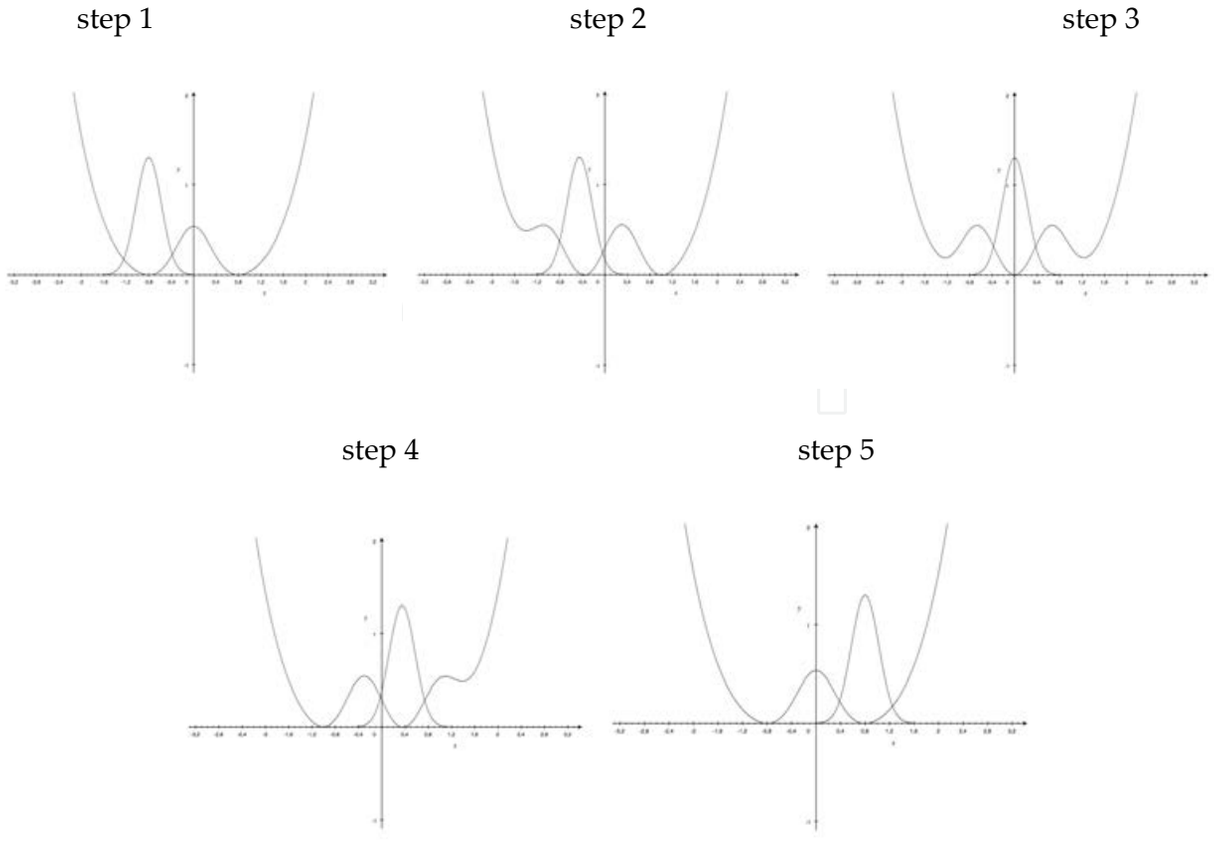


Figure 17. Potential *U (x)* + *F*. Equilibrium P (x) for the different cases. Third procedure.

manually). Note that in the non-conducting OFF state, an energy barrier is present between the metal electrodes that prevents electron transport between the electrodes (Fig. 18c). A barrier is naturally present at the interface between metal and vacuum (air) and is called work function (WF). A typical work function of stable metals is 4-5 eV. In realistic cases the barrier walls have finite slope and rounded corners due to the image force effect [19, 20]. For smaller gaps, for example when the left-hand electrode is moving towards the right-hand electrode under external field, the shape of the barrier changes, with barrier height reducing and more prominent corner rounding (Fig. 18c). Eventually, the barrier height is reduced to zero (even before the electrodes touch) that manifests the transition to the ON (close) state. The fine transient processes of Fig. 18c are often ignored in a simplified treatment, instead an abrupt transition from a high-barrier to a zero (low)-barrier state is assumed (Fig. 18d).

The bistable switches can be used to implement the three fundamental logical operations, from which all other logic functions, no matter how complex, can be derived. These operations are NOT, AND, and OR. Fig. 19 shows generic schematics for the three basic logic gates. Each logic gate consists of several distinct elements, e.g. switches and resistors. Switches can be implemented by different devices: electromechanical switches and relays, diodes, bipolar or field-effect transistors etc. For example, different Implementations of the NOT gate (inverter) are

http://dx.doi.org/10.5772/57346

shown in Figure 20. The generic switch in Figure 20a is implemented by a FET in Figures 20b and c. The resistor in NMOS implementation (Fig. 20b) can also be realized by using a transistor structure.

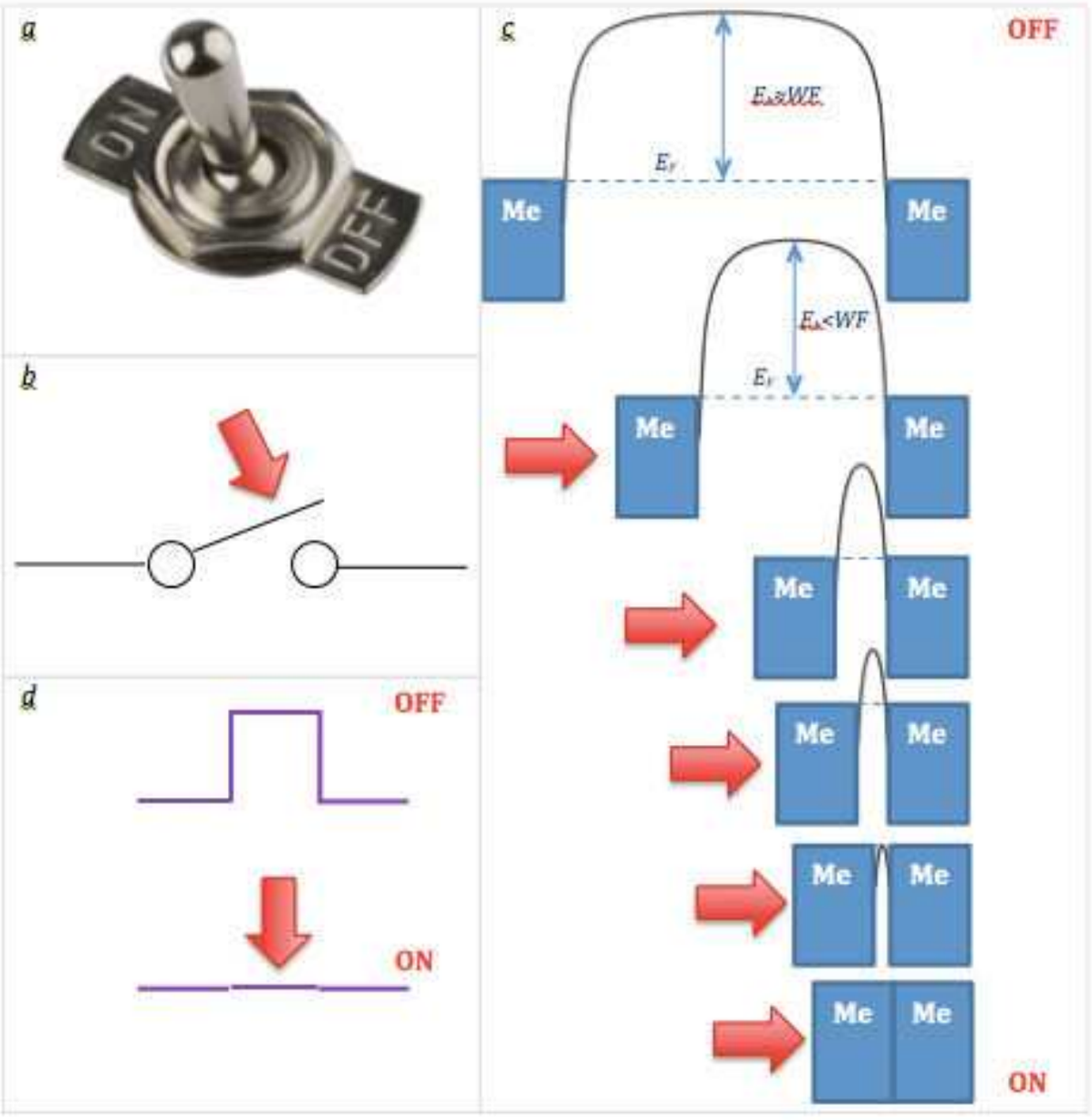


Figure 18. A bistable switch: a) An electromechanical switch example, b) Bistable switch schematics; c) Switch's barrier diagram and its evolution during OFF-ON transition; d) A simplified abrupt barrier transition model

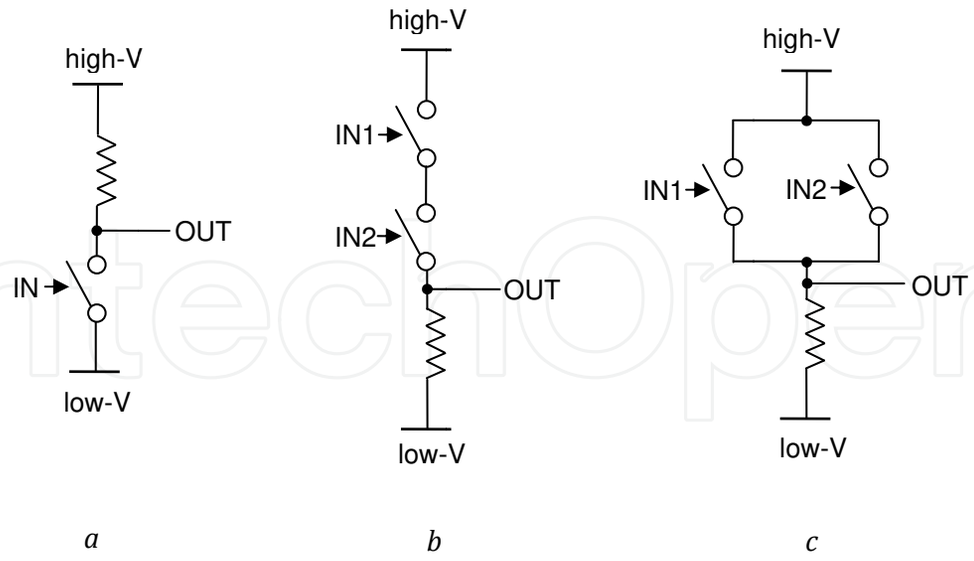


Figure 19. Generic implementations of three fundamental logic operations: (a) NOT; (b) AND, and (c) OR

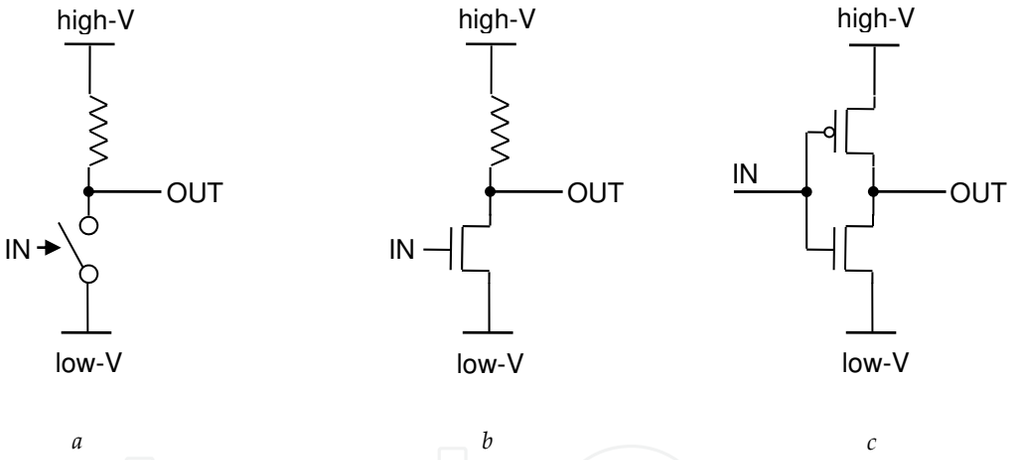


Figure 20. CMOS implementations of three fundamental logic operations: (a) NOT; (b) AND, and (c) OR

Finally in CMOS implementation of Fig. 20c, the resistor is replaced by a "complementary" FET. The role of transistors in ICT is nowadays paramount and the energy dissipation caused by these devices integrated in ICT was outlined in the previous sections.

Two basic electronic devices for information ICT will be considered here: the binary logic switch and the binary memory element. As was shown in previous sections, the *controllable barrier model* is an useful abstraction for these devices that allows for a simple and intuitive analysis of the physics-based operational limits. At least one energy barrier is always present in ICT devices, and it is fundamentally linked to the nature of information, which is a measure of *physically distinguishable states* 21. If specified locations of an information-bearing particle (e.g. electron) are used to define distinguishable states, a barrier is needed to prevent sponta-

http://dx.doi.org/10.5772/57346

neous transitions of an information–bearing particle from its 'prescribed' location (Fig. 21). The barrier must also be controllable, i.e. there must be a certain *gating* mechanism to reduce (ideally to zero) its height (or width) to allow wanted transition between informational states. Thus the three generic requirements for the implementation of a particle-based binary switch are i) the ability to detect the presence/absence of the particle in e.g., the location '0' or '1', ii) the ability to preserve on demand the particle in the location '0' or '1', and iii) the ability to move on demand the particle from '0' to '1' and from '1' to '0'.

4.1. Electronic switches

In the above, time-dependent controlled barrier transitions based on gradual adjustments of barrier shape and height were considered. In most practical cases, the treatment can be simplified assuming the barrier transitions are necessarily fast and abrupt as shown in Fig 21. Using this abrupt transition model, we offer below a quick snapshot of the current state and limitations of the electronic computing technologies due to thermal noise and quantum fluctuations. An elementary switching operation of a binary switch consists of three distinct phases shown in Fig. 21. For example, consider the switch in Fig. 21a switching from "0" to "1". The three steps are: 1) An external gating stimulus (e.g. voltage in case of charge-based devices) is applied to the barrier to reduce it from E_b to 0, 2) the particle moves from'0' to '1' location (for this transition, an additional kinetic energy E_k must be supplied to the particle), and 3) the barrier height is restored back from 0 to E_b to preserve the final '1' state. All three modes have characteristic times determined by physics and can be described by the coordinate and velocity of the information carrier/material particle, and by corresponding energies. The work required to suppress or restore the barrier is equal or larger than $E_{b.}$ It is important to note that in electronic devices, for technological reasons, this work is considered lost energy, a condition that was not present in our previous ideal model in 11.3, when the lowering or raising the barrier did not required per sè a finite amount of energy. Specifically in electric charge based devices, changes in the barrier height require changes in charge density, and as a result this always requires charging or discharging of a certain capacitor. As we have discussed above, this require a certain amount of energy dissipated. Thus, in the first order the barrier height determines the energetics of the ICT devices and it is desirable to keep E_b as low as possible for low-energy operations. How small can the energy height be? The energy barrier is needed to preserve a binary state in the presence of fluctuations, both classical (thermal noise) and quantum effect (tunneling) are present. In the following we briefly discuss these two important aspects.

Thermal noise.

The thermal noise is directly related to the fundamental result of thermodynamics, which states that each material particle at equilibrium with the environment possesses kinetic energy of ½ k_BT per degree of freedom due to thermal interactions, where k_B is the Boltzmann's constant and T is absolute temperature. The permanent supply of thermal energy to the system occurs via mechanical vibrations of atoms (phonons) and via the thermal electromagnetic field of photons (background radiation). Thus the barrier height, E_b, must be large enough to prevent spontaneous transitions (errors) [22] that occur when the particle spontaneously acquires

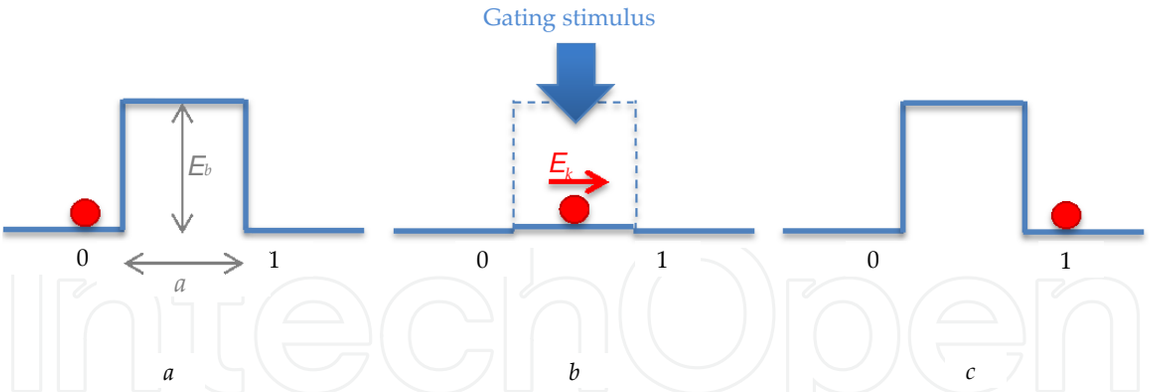


Figure 21. Three phases of abrupt binary switching

thermal energy large enough to jump over the barrier. This can easily happen if the kinetic energy of the particle E is larger than the barrier height E_b. This can be easily seen, by estimating the probability for over-barrier transition from the Boltzmann distribution:

$$f(E) = A\exp\left(-\frac{E_b}{k_B T}\right) \tag{23}$$

The probability of over-barrier transition is equivalent to the probability that the particle gains energy $E>E_b$ which probability is obtained by integration of (23):

$$p = \int_{E_b}^{\infty} f(E)\, dE = A\int_{E_b}^{\infty} \exp\left(-\frac{E}{k_B T}\right) dE = Ak_B T \exp\left(-\frac{E_b}{k_B T}\right) \tag{24}$$

The coefficient A can be found from the normalization condition for (23):

$$\begin{aligned} 1 &= \int_{0}^{\infty} f(E)\, dE = A\int_{0}^{\infty} \exp\left(-\frac{E}{k_B T}\right) dE = Ak_B T = 1 \\ A &= \frac{1}{k_B T} \end{aligned} \tag{25}$$

Substituting (25) into (24) obtain that the probability of over-barrier transition is

$$p_{o-b} = \exp\left(-\frac{E_b}{k_B T}\right) \tag{26}$$

The minimum barrier height can be found from the distinguishability condition, which requires that the probability of errors $p < 0.5$, in which case the switch is being operated at the

http://dx.doi.org/10.5772/57346

threshold of distinguishability. Solving (24) for p = 0.5, obtain the Boltzmann's limit for the minimum barrier height, E_{bmin} as

$$E_{b\min} = k_B T \ln 2 \approx 0.7 k_B T \sim k_B T \tag{27}$$

Of course error probabilities much less than 0.5 are required in practice, and therefore the barrier height E_b must be larger. For example in modern DRAM the probability of one erroneous bit is ~10^{-9} in a month [10].

The barrier model along with (23) can be further applied to derive the classic formula for the thermal (Nyquist-Johnson) noise, which plays a fundamental role in analog devices:

$$\left\langle V_n^2 \right\rangle = 4 k_B T R \Delta f \tag{28}$$

($\left\langle V_n^2 \right\rangle$is the variance of the noise voltage across a resistor due to thermal agitations, R is the resistance and Δf is operational bandwidth. A derivation of (28) using the barrier model is considered in [23]).

Quantum effects

Another class of errors that impose limits on device scaling are quantum errors, which occur due to quantum mechanical effects. These effects play a measurable role in a system whose energy (*E)*, momentum (*p*), space (*l*) and time (*t*) are such that the characteristic physical parameter, the action, $S \sim E\,t \sim p\,l$, is comparable to the *quantum of action h*=6.63×10^{-34} J s (Planck's constant). The corresponding relations are known as *Heisenberg Uncertainty Principle*:

$$\Delta x \cdot \Delta p \sim \frac{h}{2} \tag{29}$$

From (29), the minimum size of a scaled computational element or switch (Fig. 21) is

$$L_{\min} > \Delta x \sim \frac{h}{2\Delta p} = \frac{h}{2\sqrt{2mE_b}} \tag{30}$$

where m is the mass of the information-bearing particle, for example that of the electron.

The Heisenberg relation (29) and its derivative (30) can be used for an elementary derivation of an analytical form of tunnelling probability (known as Wentzel-Kramers-Brillouin (WKB) approximation):

$$p_{WKB} \sim \exp\left(-\frac{2\sqrt{2m}}{\mathrm{h}} \cdot a \cdot \sqrt{E_b} \right) \tag{31}$$

Note that (30) and (31) emphasizes the parameters controlling the tunnelling process. They are the barrier height E_b and barrier width a as well as the mass m of the information-bearing particle. If separation between two wells is less than L_{min} (30) the barrier structure of Fig. 21a would allow significant tunnelling, which will destroy the binary information. Also, parasitic leakage current will considerably increase the total power consumption. For a numerical example using E_b=0.1eV and the effective mass of electron in semiconductor m^*=0.19m_e (the transverse electron effective mass in Si) obtain from (30)

$$L_{\min} \sim \frac{6.63 \cdot 10^{-34}}{2\sqrt{2 \cdot 0.19 \cdot 9.11 \cdot 10^{-31} \cdot 0.5 \cdot 1.6 \cdot 10^{-19}}} \approx 4.5nm,$$

which is an approximate minimum channel length of the Si *logic* FET[21]. This assessment is consistent with ITRS, which projects the minimal physical gate length in logic FET to be ~5 nm [10]. At this scale, leakage due to quantum mechanical tunnelling will be very significant and may limit usage of these 'ultimate' devices in many practical applications.

4.2. Memory elements

Next consider ultimate dimensional scaling of the memory elements. To estimate the needed barrier properties for memory, one needs to understand the limits on electrical conductance, which can be done using another form of the Heisenberg relations [24]:

$$\Delta E \Delta t = \frac{h}{2} \tag{32}$$

Let's consider an elementary act of electrical conductance for an electron passing from reservoir **A** with energy E_A to reservoir **B** with energy E_B (Fig. 22).

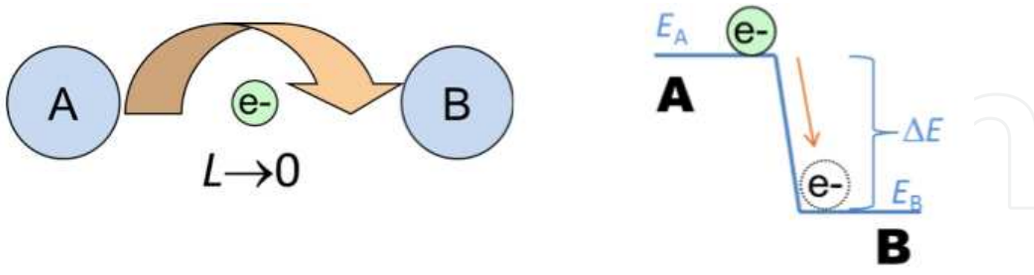


Figure 22. Illustration to the derivation of quantum conductance

The corresponding voltage (potential difference) between **A** and **B,** V_{AB} and the current, $I_{AB,}$ flowing from **A** to **B** are:

$$|V_{AB}| = \frac{E_A - E_B}{e} = \frac{|\Delta E|}{e} \tag{33}$$

http://dx.doi.org/10.5772/57346

$$I_{AB} = \frac{e}{\Delta t} \tag{34}$$

The minimum passage time Δt from (32) is:

$$\Delta t = \frac{h}{2\Delta E} = \frac{h}{2eV} \tag{35}$$

Putting (35) into (34), and taking into account Ohm's law, i.e. $I=V/R$, obtain:

$$I_{AB} = \frac{2e^2}{h} \cdot V = \frac{V}{R_0} \tag{36}$$

where

$$R_0 = \frac{h}{2e^2} = 12.9k\Omega \tag{37}$$

is quantum resistance. A related parameter is quantum conductance:

$$G_0 = \frac{1}{R_0} = \frac{2e^2}{h} \tag{38}$$

The quantum resistance/conductance sets the limit on electrical conductance in a one-electron channel *in the absence of barriers*.

$$I_0 = \frac{V}{R_0} = \frac{V}{12.9k\Omega} \tag{39}$$

If a barrier is present in the electron transport system, the conductance will be decreased due to the barrier transmission probability $p_T < 1$. The electrical conductance in the presence of barrier is obtained by multiplying the barrier-less quantum conductance (38) by the barrier transmission probability:

$$G = \frac{1}{R} = G_0 \cdot p_T \tag{40}$$

Eq. (40) is a form of the *Landauer formula* [25] for a one-electron conductive channel.

Let now consider as an example the insulator-conductor-insulator memory element shown in Fig. 23, which is representative of floating gate cell used in flash memory. In memory cells that store electron charge, two distinguishable states 0 and 1 are created by the presence (e.g. state 0) or absence (e. g. state 1) of electrons in the charge storage node. In order to prevent losses of the stored charge, the storage node is defined by energy barriers of sufficient height E_b to retain charge (Fig. 23). Assume only one electron is stored. The store time (or corresponding characteristic escape time) is:

$$t_s = \frac{e}{I_{leak}} \tag{41}$$

The two mechanisms of the charge loss are over-barrier leakage and through-barrier tunnel leakage. In both cases the leakage current from the storage node can be calculated from the Landauer formula (40):

$$I_{leak} = G_0 \cdot V \cdot p_T \tag{42}$$

In the following, the thermal voltage $V = \frac{k_B T}{2e}$ will be used as a lower bond.

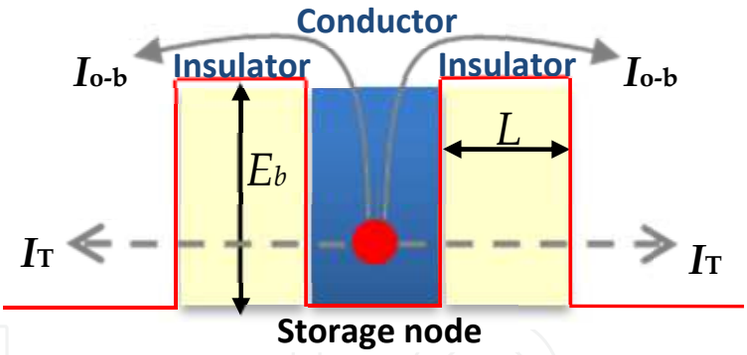


Figure 23. An insulator-conductor-insulator memory element, representative of flash memory.

The probability of thermal over-barrier transitions is the Boltzmann probability - From (42) and (26) the one-electron over-barrier current $I_{o\text{-}b}$ is:

$$I_{o-b} = 2\frac{e}{h} \cdot k_B T \cdot \exp\left(-\frac{E_b}{k_B T}\right) \tag{43}$$

The factor of 2 in (43) appears because escape is possible over either of two barriers that confine an electron as shown in Fig.23.

The electron escape time (the retention time) due to over-barrier transport is:

http://dx.doi.org/10.5772/57346

$$t_{o-b} = \frac{h}{2k_B T} \exp\left(\frac{E_b}{k_B T}\right) \tag{44}$$

If over-barrier leakage is the only mechanism of charge loss (when the barrier width a is sufficient to suppress tunneling), the escape time is equal to the one-electron retention time, $t_{o\text{-}b}=t_s$.

For a specified t_r, the required minimum barrier height is:

$$E_{b\min} = k_B T \ln\left(\frac{2k_B T}{h} t_s\right) \tag{45}$$

In the case of the 'minimum nonvolatile memory' requirement, i.e. t_r>10 years, (45) gives E_{bmin} ≥ 1.29 eV (~50k_BT) at T=300 K.

A second source of charge loss is electron tunneling. The corresponding tunneling current I_T is:

$$I_T = 2\frac{e}{h} \cdot k_B T \cdot \exp\left(-\frac{2\sqrt{2m}}{\mathsf{h}} \cdot a \cdot \sqrt{E_b}\right) \tag{46}$$

The electron escape time due to tunneling is:

$$t_T = \frac{h}{2k_B T} \exp\left(\frac{2\sqrt{2m}}{\mathsf{h}} \cdot a \cdot \sqrt{E_b}\right) \tag{47}$$

The total retention time due to both mechanisms can be estimated as:

$$t_r = \frac{e}{I_{o-b} + I_T} \tag{48}$$

Suppose that the barrier height is large enough to suppress over-barrier escape, i.e. E_b>>E_{bmin}, where E_{bmin} is given by (45). In this case, the store time will be determined by the tunneling time, t_T: t_s≈ t_T. The minimum barrier width for a specified store time, can be estimated from (47), e.g. for t_s=10 years:

$$a_{\min} = \frac{\mathsf{h}}{2\sqrt{2mE_b}} \ln\frac{k_B T}{h} t_s \tag{49}$$

As a numerical estimate for t_s>10 years, $E_{bmin} \geq 1.29$ eV, $m=m_e$ and T=300 K, (48) gives $a_{\min} \sim 5$ nm.

As follows from the above, in order to obtain a nonvolatile electronic memory cell, sufficiently high barriers must be created to retain the charge for a long period of time. If different practical factors are taken into account (such higher temperature e.g. T=400 K, lower effective electron mass in solids, e.g. $m^*= 0.5\ m_{0,}$ many-electron distribution in solids etc,. for >10 y retention, the minimal barrier height E_b is ~2 eV (~77 k_BT), and thickness a>5 nm [26]. The corresponding practical minimum size of the floating gate cell is ~10 nm [26]. Large barriers also result in high voltages required for memory operation: ~5 V for READ and ~15 V for WRITE [26].

4.3. Energy per bit operation

As it was mentioned earlier, in charge-based devices, changes in the barrier height require changes in charge density, and as a result this always requires charging or discharging of a certain capacitor associated with the device (e.g. a gate capacitor C_g in the case of FET, interconnect line capacitance C_{line} etc.). It was shown in section 11.2 that when a capacitor C is charged from a constant voltage power supply, the energy of $CV^2/2$ is dissipated, and operation of binary devices in this regime is sometimes referred as *irreversible* switching. The total energy per bit operation depends on the device barrier height E_b and the number of electrons N_e involved in the switching process. The minimum energy needed to suppress the barrier (e.g. by charging the gate capacitor) is equal to the barrier height E_b. Restoration of the barrier (e.g. by discharging gate capacitance) also requires a minimum energy expenditure of E_b. Thus the minimum energy required for a full switching cycle is at least $2E_b$. Additional kinetic energy E_k (typically ~E_b) also needs to be supplied to electrons to enable the transition, If N_e is the number of electrons involved in the switching transition between two wells, the total minimum switching energy is

$$E_{bit\min} = 2E_b + N_e E_k = \left(N_e + 2\right)k_B T. \tag{50}$$

If N_e=1,

$$E_{bit\min} = 3k_B T \approx 10^{-20} J, \tag{51}$$

and this is a lower boundary for a logic operation. For a nonvolatile memory device, that requires a minimum barrier height $E_b \sim 50 k_B T$ the minimal energy to store one electron is $E_{bit} \sim 150\ k_B T$.

The above analysis considered individual logic and memory devices operating in a single electron limit. In practice, a larger number of electrons, N_{el}, is needed to support communication between different devices in the system. For logic operations, one 'upstream' binary switch controls/communicates with several 'downstream' binary switches. The number of the downstream devices that are driven by a given upstream device is called 'fan-out' (FO). A typical fan-out in the baseline microprocessors is four (FO4). For communication, the devices

http://dx.doi.org/10.5772/57346

are interconnected with metal wires, and in a 2D layout at least one electron needs to be sent to each of the 'downstream' gates, thus, at least four electrons needs to be provided by the 'upstream' devices, and according to (50) $E_{bit} > 6\ k_B T$. In practice, the number of electrons is much larger to ensure communication reliability. Next, at least a few long interconnects are needed to ensure communication between the information processing system and the outside world (e.g. I/O). The energy costs associated with long interconnects can be estimated as energy needed to charge/discharge a metal line of length L:

$$E \sim C_{line}\ V^2 \tag{52}$$

Using line capacitance of $C_{line} \sim \varepsilon_0 L$ and nearly minimal distinguishable voltage: $V \sim k_B T/e$ obtain as a lower boundary for the communication energy per unit length:

$$E > \varepsilon_0 \left(\frac{k_B T}{e} \right)^2 \sim 10^{-14} \frac{J}{bit \cdot m} \tag{53}$$

For an example of a long wire along a 1 cm chip the limiting communication energy is 10^{-16} J/bit, i.e. 10,000x times more than the minimal energy required for computation!

The long wire considerations are critical for memory that typically is organized in regular X-Y arrays of memory cells. In many instances the properties of interconnecting array wires determine the operational characteristics of the memory system. A given cell in an array is selected (e.g. for read operation) by applying appropriate signals to both interconnect lines, thus charging them. The relatively large operating voltage of flash results in rather large line charging energy. For a memory cell pitch of 10nm and a 128×128 array the line capacitance is ~10^{-14} F [27].For write operation with V_{write}~15V, the write energy is ~10^{-12} J/line. (In practical flash memory devices the read energy is of the order of 10^{-13}-10^{-11} J/bit read and 10^{-9}-10^{-10} J/bit write [28, 29]).To summarize, electrons flowing in metal wires constitutes the main component in energy consumption in the electron based devices. It can also be argued that fixed wiring is among the main factors limiting the efficiency of computational systems. Energy consumption (per bit) in different components of ICT as discussed in this chapter is summarized in Table 2.

5. Open issues and conclusions

Even with steady decreases in the energy required to switch a bit as shown in Fig. 4, it appears that the 'effective' energy required to switch a bit is decreasing at a slower pace. The other essential components of an information processing system are assuming a relatively more significant role in system energy consumption. For example, increases in energy utilization by I/O systems, increases in memory access energy costs due to the array structure of the memory architectures, increases in power consumption by the internal chip wires, device leakage in the OFF state, etc., are consuming a greater share of information processing system energy.

	Fundamental limit	Baseline technology
	Logic	
Barrier height	$k_BT$ln2=18meV= 3×10^{-21}J	10^{-19}J ~ 0.5-1eV~ 24k_BT
Logic device	~3k_BT ≈ 80meV ≈ 10^{-20} J	3×10^{-17}J* ≈ 188eV ≈ 7,250k_BT
Logic circuit	>6k_BT ≈ 160meV ≈ 2×10^{-20}J	10^{-16}J* ≈ 625eV ≈ 24,150k_BT
	Nonvolatile Memory	
Barrier height	~50k_BT~1.3eV~2×10^{-19}J	~77k$_B$T ~ 3×10^{-19}J
Memory device	~150k_BT~4eV~6×10^{-19}J	~230 k$_B$T ~ 10^{-18} J
Memory array	2×10^5 k_BT~10^{-15}J	10^{-11}-10^{-13}J
I/O	10^{-16} J	10^{-11} J

*see Table 1

Table 2. Energy consumption (per bit) in ICT: A summary

Leading edge devices today utilize slightly over three orders of magnitude more energy than the k_BT ln2 thermodynamic limit. If current trends continue, it is likely that further reduction in the energy per bit of a device will not be accompanied by corresponding decreases in effective energy-per-bit when viewed at the system level. Moreover, a second issue associated with continuing scaling of device features and supply voltages is that thermal and tunneling noise will require increased use of error correction mechanisms.

It has been observed that it might be possible to operate a switch at energy close to k_BT ln2 for irreversible switching procedures and even lower for entropy preserving switching procedures. This possibility was examined and shown to be theoretically possible; however, the side attributes associated with achieving such a functional device may not be acceptable in practice. For example, there is a need for very slow operation of the device that may be untenable and the energy recovery mechanisms associated with energy storage and retrieval are difficult to implement without incurring energy loss. However, even if one assumes that this can be achieved without any overhead penalties, the communication and fan-out cost of interconnects and I/Os may make this achievement almost invisible. It is also unlikely that energetics of memory devices can be significantly changed. These limitations are even more apparent in charge based devices where *the main source of energy consumption is due to electrical charging large capacitances in metal wires*.

Having said all that, extremely low energy computation may be achievable in systems nased on fifferent technologies. One example is represented by living systems, where it has now been established that individual cells, the smallest units of living matter, possess amazing computational capabilities, and are indeed the smallest known information processors [30, 31]. As argued in a number of studies, individual living cells, e.g. bacteria, have the attributes of a Turing Machine, capable of a general-purpose computation [30, 32, 33], and von Neumann's Universal Constructor, i.e. *computer making computers*[32] (DNA molecule acts as nonvolatile memory of the cell computer, while many proteins in cell's cytoplasm have as their primary function the transfer and processing of information, and are therefore can be regarded as logical elements of the biological cell processor [34, 35, 36, 37]). The Universal Constructor

http://dx.doi.org/10.5772/57346

model is a useful concept for the estimation of the information content of a living cell, for example for the *E.coli* bacterium the estimated information content is ~10^{11}-10^{12} bit [11, 23] (interestingly, experimental entropy reduction measurements of the informational content of bacterial cells using microcalorimetric techniques yielded very similar results [38]). Assuming a conservative edge of cell's information content, which is 10^{11} bit and ~ 3000 s for reproduction time of a bacterial cell obtain ~10^{7} equivalent bits that must be processed per second (equivalent binary throughput). The power consumption of *E.coli* is about 1.4×10^{-13} W so that from (11) the energy per equivalent binary operation in the cell can be calculated to be ~10^{-20} J or <10 k_BT. Note, that this is the total energy per bit, taking into account logic, memory, I/O (e.g. sensing, ribosomial synthesis etc.).

	Biological Cell Processor	**Baseline ICT**
Memory	10^7 bit	10^7
Logic	10^6 bit	10^6
Energy per bit	~10^{-20} *J*(average system level)	10^{-13} (Memory) 10^{-16} (Logic)
Binary throughput	10^7 bit/s	10^7 bit/s
Task time	3000 s	3000 s
Total energy per task	10^{-10} J	10^{-6} J

Table 3. Comparison of the two technologies

The estimated energy utilization per switching event is quite impressive. It can be compared to an equivalent electronic system consisting of the same number of logic and memory elements implemented in baseline technology (Table 2). A comparison of the two technologies reveals that the biological cell processor operates with the four orders of magnitude lower energy than the baseline electronic processor (Table 3).

What makes biological cell a superior information processor relative to the performance of ultimately scaled semiconductor technology? It appears that several simple physics based arguments can be made:

1. Heavier mass of information carrier allows for denser logic and memory. As was argued in section 11.4, a heavier mass for the information carrier allows for smaller separation between distinguishable states and therefore more devices/states per unit volume or area. According to (30), a heavier mass results in smaller device size. For example, DNA memory uses molecular fragments (nucleotides) as information carriers, each consisting of more than 10 atoms. The molecular information carriers are densely packed in a linear array with distance between nucleotides of only 0.34 nm. By comparison, the minimum size of an electron memory cell is ~10 nm. This dimensional difference gives insight into the 1000x difference in volumetric memory density between electronic and DNA memory, i.e. 10^{16} bit/cm^3 of electronic memory vs. 10^{19} bit/cm^3 for DNA memory. Also, protein logic devices consist of arrangements of many atoms, resulting in total device size of ~5 nm or less, which can explain about 10x higher protein logic density compared to ultimately scaled transistors. In fact, a vision of cellular enzyme proteins as conformon-based "soft-

state nanotransistors" has recently been recently introduced in a book by Ji and contrasted with electron-based solid-state transistors [39].

2. Utilization of ambient thermal energy allows for energy minimization in logic circuits. For semiconductor systems thermal energy (~k_BT) must be managed as it may destroy the state or divert the information carrier from its intended trajectory; for example in communication between several logic elements. In order to overcome the deleterious effects of thermal energy each logic element must contain a barrier $E_b > k_BT$. Moreover, in communication with other elements, N carriers must be sent to the recipient elements, each of which must have kinetic energy $Ek > k_BT$. As result the total energy per bit operation, as it was derived in section 11.5, becomes (50):

$$E_{bit\,\min} = 2E_b + N_e E_k = (N_e + 2)k_B T$$

and it can be significantly large, usually >1000 k_BT.

In contrast, biomolecular computing systems utilize thermal energy to effect data exchange/transmission between e.g. logic-to-logic or memory-to-logic elements. All computational molecules move within the cell's volume by thermally excited quasi-random walk with almost no extra energy required, thus E_k~k_BT and the second term is minimized. Biological systems actually use thermal energy in the transmission of information and in the realization of work-related tasks40. Examples of beneficial use of non-equilibrium fluctuations are present also at micro and nano scale level: see e.g. the paradigmatic phenomenon of Stochastic Resonance41.

3. Flexible/on-demand 3D connections/routing allow for minimization of the communication carriers. Referring to (50), in silicon systems most energy is consumed by interconnect. This is due to the need to pump a large number, N, of carriers (electrons) into the interconnecting wire for reliable communications. As was argued in section 11.5, for reliable communication, N must dramatically increase for longer path lengths and more receiving devices (fan out). In electrical circuits the connection paths are pre-determined in 2-D networks, and in many instances, the electron travels a long distance. A problem of electrical interconnects is the statistical behavior of discrete charges, in other words electrons are free to move along the line. Therefore a large number of electrons is needed for reliable branched communication to reduce thermal and shot noise. Electrons flowing in 2-D networks of metal wires constitute the main component in energy consumption in the electron based systems. In contrast, 'devices' in cells (e.g. proteins or RNA) are usually free to travel in all three dimensions within the cell and they don't follow a fixed path. Due to the shape-specific molecular recognition (e.g., lock-and-key interactions) a 'deterministic' or 'point-to-point' communication of information packages within the processor is obtained, with smaller number of carriers, resulting on lower energy.

4. Array-free organization of DNA memory enables the minimization of the energy for memory access. A core system-level challenge resulting in the excessive energy consumption in the silicon microcomputer is that memory access to support computations takes too much energy. Organizing solid-state memory in cross-bar arrays, while an elegant

http://dx.doi.org/10.5772/57346

solution at larger scale, contributes to excessive energy dissipation due to line charging during memory access as given by:

$$E \sim C_{line} V^2$$

The long wires needed to connect memory elements in an array result in a large line capacitance C_{line}, which together with a large access voltage required for nonvolatile electron-based memory, yielding 10^{-13}-10^{-11} J per randomly accessed bit. In contrast, the DNA memory in the cell uses array-less organization that can be viewed as similar to access to tape or hard disc drive. Multiple read heads (formed by RNA polymerase protein) are used for independent simultaneous access to different parts of the DNA memory, thus this is a highly parallel process.

5. Hybrid digital & analog information processing. As it is argued in [36] the cell processor is a hybrid state machine operating in both digital and analog modes. For example, DNA memory is a digital unit while the sensory information the cell receives from its environment is mostly analog. The protein-based computing often represents and processes information in analog form, with state variables encoded in concentrations of protein molecules.

From the above discussion, it would appear that the performance and energy efficiency of the general purpose electronic ICT so widely prevalent today is becoming increasingly difficult to improve within the context of its implementation in semiconductor technology. In order to further improve performance and energy efficiency in computation, it may be necessary to invent a new general-purpose architecture and/or implementation technology. The living cell, whose dimensions are only on the order of a few microns, is a powerful information processor that utilizes extremely small amounts of energy (*~10 kT* per bit) and achieves high functional performance. It may be that inspiration can be drawn from the architecture and technologies used by the cell to develop future information processing systems. The cell is a very complex system about which much is yet to be learned but it may provide suggestions for a pathway for more energy-efficient information processing.

Author details

Victor Zhirnov[1], Ralph Cavin[1] and Luca Gammaitoni[2]

1 Semiconductor Research Corporation, USA

2 NiPS Laboratory, Università di Perugia, Italy

References

[1] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices* (1998 Cambridge University Press)

[2] K. Bernstein, R.K. Cavin, W. Porod, A. Seabaugh, J. Welser, "Device and architecture outlook for beyond CMOS switches, *Proc. IEEE* 98 (2010) 2169-2184

[3] J. J. Welser, G. I. Bourianoff, V. V. Zhirnov, R. K. Cavin, "The quest for the next information processing technology", J. Nanoparticle Res. 10 (2008) 1-10

[4] S. Carnot, Reflections on the Motive Power of Heat and on Machines fitted develop this Power, 1824, Translated by R. H. Thruston, ASME (1943).

[5] S. Shankar, R. K. Cavin, and V. V. Zhirnov, "Computation from Devices to System-Level Thermodynamics," *Electrochem. Soc. Transactions* 25 (2009) 421-431

[6] J. von Neumann, Fourth University of Illinois lecture, in *Theory of Self-Reproducing Automata*, A.W. Burks, ed., p. 66. Univ. of Illinois Press, Urbana (1966)

[7] R. Landauer, "Irreversibility and heat generation in the computing process", *IBM J. Res. Dev.* 5 (1961) 183-191

[8] C. H. Bennett, "The thermodynamics of computation - a review", *Int. J. Theoretical Physics* 21 (1982) 905-940

[9] B. Nordman, S. Lanzisera, "Electronics and network energy use: Status and prospects", 2011 IEEE Intern. Conf. Consumer Electron. , pp 245-246

[10] International Technology Roadmap for Semiconductors (ITRS), www.itrs.net

[11] R. K. Cavin, P. Lugli, V. V. Zhirnov, "Science and Engineering Beyond Moore's Law", *Proc. IEEE* 100 (2012) 1720

[12] T. Sekiguchi, K. Ono, A. Kotabe, Y. Yanagawa, "1-Tbyte/s 1-Gbit DRAM architecture using 3-D interconnect for high-throughput computing", IEEE J. Solid-State Circ. 46 (2011) 828-837

[13] W. Gitt, "Information - the 3rd fundamental quantity", *Siemens Review* 56 (1989) 36-41

[14] H. Moravec, "When will computer hardware match the human brain?" *J. Evolution and Technol.* 1 (1998) 1-12

[15] D. E. Dillenberger, D. Gil, S. V. Nitta, M. B. Ritter, "Frontiers of information technology", IBM J. Res. & Dev. 55 (2011) 1:1 - 1:13

[16] Luca Gammaitoni (2012): There's plenty of energy at the bottom (micro and nano scale nonlinear noise harvesting), Contemporary Physics, 53:2, 119-135

[17] C. E. Shannon, A Mathematical Theory of Communication, The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948

[18] L. Gammaitoni, 2011, arXiv:1111.2937; L. Gammaitoni, Nanoenergy Letters, 5, 10, 2013..

[19] J. G. Simmons, "Generalized formula for electric tunnel effect between similar electrodes separated by a thin insulating film", *J. Appl. Phys*. 34 (1963) 1793

[20] L. Gammaitoni, D. Chiuchiù, work in preparation, 2014.

[21] R. U. Ayres, *Information, Entropy and Progress*. New York : AIP Press, 1994.

[22] L. Gammaitoni, "Noise limited computational speed", Applied Physics Letters, 11/2007, Volume 91, p.3, (2007)

[23] V. V. Zhirnov and R. K. Cavin, *Microsystems for Bioelectronics* (Elsevier 2010)

[24] I. P. Batra, "Origin of conductance quantization", Surf. Sci. 395 (1998) 43-45

[25] Y. Imry and R. Landauer, "Conductance viewed as transmission", Rev. Mod. Phys. 71 (1999) S306-S312

[26] V. V. Zhirnov and T. Mikolajick, "Flash Memories", in: "Nanoelectronics and Information Technology", by R. Waser (Ed.) Wiley-VCH 2012, pp. 623-634

[27] V. V. Zhirnov, R. K. Cavin, S. Menzel, E. Linn, S. Schmelzer, D. Bräuhaus, C. Schindler and R. Waser, "Memory Devices: Energy-Space-Time Trade-offs", *Proc. IEEE* 98 (2010) 2185-2200

[28] L. M. Grupp, A. M. Caulfield, J. Coburn, S. Swanson, E. Yaakobi, P. H. Siegel, J. K. Wolf "Characterizing Flash Memory: Anomalies, Observations, and Applications", *MICRO'09*, Dec. 12-16, 2009, New York, NY, USA, p.24-33

[29] N. Derhacobian, S. C. Hollmer, N. Gilbert, M. N. Kozicki, "Power and energy perspectives of nonvolatile memory technologies", Proc. IEEE 98 (2010) 283-298

[30] S. Ji, "The cell as the smallest DNA-based molecular computer", Biosystems 52 (1999) 123-133

[31] R. Sarpeshkar, Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications, and Bio-Inspired Systems (Cambridge University Press 2010)

[32] A. Danchin, "Bacteria as computer making computers", *FEMS Microbiol. Rev*. 33 (2009) 3-26

[33] C. T. Fernando, A. M. L. Liekens, L. E. H. Bingle, C. Beck, T. Lenser, D. J. Stekel, J. E. Rowe, "Molecular circuits for associative learning in single-celled organisms", *J. R. Soc. Interface* 6 (2009) 463-469

[34] D. Bray, "Protein molecules as computational elements in living cells", *Nature* 376 (1995) 307-312

[35] N. Ramakrishnan, U. S. Bhalla, and J. J. Tyson, "Computing with proteins", *Computer* 42 (2009) 47-56

[36] L. F. Agnati, D. Guidolin, C. Carone, M. Dam, S. Genedani, K. Fuxe, "Understanding neuronal cellular network architecture", *Brain Res. Rev.* 58 (2008) 379-399

[37] A. Wagner, "From bit to it: How a complex metabolic network transforms information into living matter", *BMC Systems Biology* 1 (2007) 33

[38] W. W. Forrest, "Entropy of microbial growth", *Nature* 225 (1970) 1165-1166

[39] Sungchul Ji, Molecular Theory of the Living Cell: Concepts, Molecular Mechanisms, and Biomedical Applications, Springer, 2012.

[40] P. Hanggi, F. Marchesoni, "Artificial Brownian motors: Controlling transport on the nanoscale", Rev. Mod. Phys. 81 (2009) 387-442

[41] L. Gammaitoni, P. Hanggi, P. Jung, F. Marchesoni, "Stochastic resonance", Rev. Mod. Phys. 70 (1998) 223-287