BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 58, No. 1, 2010
DOI: 10.2478/v10175-010-0012-8

Heat dissipation and temperature distribution in long interconnect lines

K. GNIDZINSKA[1]*, G. DE MEY[2], and A. NAPIERALSKI[1]

[1] Department of Microelectronics and Computer Science, Technical University of Lodz, 221/223 Wolczanska St., 90-924 Lodz, Poland
[2] Department of Electronics and Information Systems, University of Ghent, Sint-Pietersnieuwstraat 41, B-9000, Belgium

Abstract. Thermal and time delay aspects of long interconnect lines have been investigated. To design a modern integrated circuit we need to focus on very long global interconnects in order to achieve the desired frequency and signal synchronization. The long interconnection lines introduce significant time delays and heat generation in the driver transistors. Introducing buffers helps to spread the heat production more homogenously along the line but consumes extra power and chip area. To ensure the functionality of the circuit, it is compulsory to give priority to the time delay aspect and then the optimized solution is found by making the power dissipation as homogenous as possible and consequently the temperature distribution T (relative to ambient) as low as possible. The technology used for simulations is 65 nm node. The occurring phenomena have been described in a quantitative and qualitative way.

Key words: interconnect lines, heat dissipation, delay, temperature distribution.

1. Introduction

The CMOS technology is undergoing continuous improvements concerning the used materials, the physical design and the sizes of the elements. The latter is called scaling and in fact concerns mostly the transistor's sizes. As a result the characteristic size lessens, the density of the circuit increases and the chip area remains unchanged, and consequently new problems occur.

We would like to focus on global interconnection lines. In more advanced (smaller) technologies (so called nanometric or deep submicron) we experience a brand new range of difficulties. A typical approach to the matter is limited to time delay caused by the interconnect and solutions to this phenomenon. We would like to concentrate more on the thermal point of view and also taking into account time delay aspect. Even though advanced technologies means smaller and smaller transistors, interconnect lines, especially the global ones, are not getting shorter. The scale of integration is rising, the number of transistors on one chip is reaching 10^7. This is why the length of global interconnects is not scaling down along with the whole technological progress. The problem is that if the length of the interconnect is not changing and all the components are getting smaller, it is more and more difficult to drive for example a 1 cm long path by a single inverter. One of the possibilities is to change the materials. For example to use ILD – interlevel dielectrics to reduce interconnect's capacitance by decreasing dielectric constant of the layers below [1]. This technique is used in the theoretical model of 65 nm-technology.

A possible approach to the problem of long interconnection lines is to introduce buffers along the line. Putting an inverter in the middle of it splits the line into two parts driven by different inverters. As energy loss (in other words heat production) on an inverter is proportional to its load and parasitic capacitance, the heat dissipation is more spread along the line. However another important aspect is that placing buffers influences time delay depending on the aspect ratio of the transistors the buffers are made from. Using wider buffers can lead to a decrease of the delay but also gives a rise to an increase of the power dissipation. The increase of the power dissipation in a relatively small area induces higher temperature rises. There is a need to find an optimum solution, taking into account which features of the circuit are of higher importance.

2. Heat generation in CMOS

As a CMOS inverter is switched, the current flows through the series connection of the NMOS and PMOS transistors between power supply and ground. In CMOS technology the major power dissipation occurs during switching. In fact an inverter has its intrinsic, parasitic capacitance C, which is charged or discharged by the current while switching. The power dissipation (for half a period – one charging or discharging) can be calculated as follows:

$$E = \frac{1}{2} C V_{DD}^2. \tag{1}$$

If there is any load capacitance (C_{load}), its effect on the power dissipation is exactly the same (Fig. 1). While switching an inverter has to reload not only its own parasitic capacitance but also the load capacitance. Capacitances make a parallel connection and total heat production in the inverter rises to:

$$E = \frac{1}{2} (C + C_{load}) V_{DD}^2. \tag{2}$$

*e-mail: kmg@dmcs.p.lodz.pl

Therefore we can notice that depending on proportion between C and C_{load} the load can have a significant influence on heat production in an inverter.

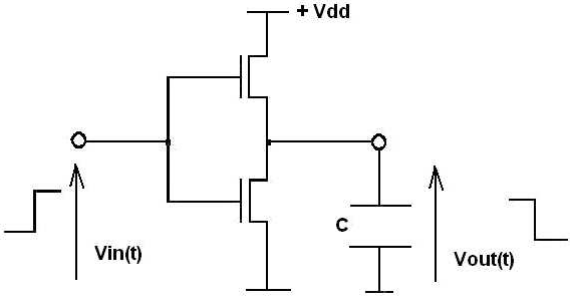

Fig. 1. Inverter with parasitic capacitance

CMOS technology is said to have no static power dissipation. In reality, the effects of the non-zero quiescent current existence can be neglected in case of so called long-channel devices. In more advanced technologies, this assumption is no longer valid. The border length of the channel is said to be 0.25 μm. Below this point a new set of phenomena is observed. In this case the most important is the subthreshold conduction, also known as the weak-inversion conduction. The effect is perfectly noticeable when we compare I_D versus V_{GS} curves in logarithmic scale of 65 nm and 0.35 μm technologies (Fig. 2 and Fig. 3).

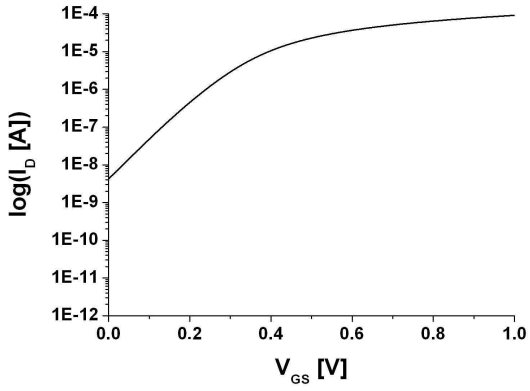

Fig. 2. I_D versus V_{GS} in logarithmic scale – 65 nm

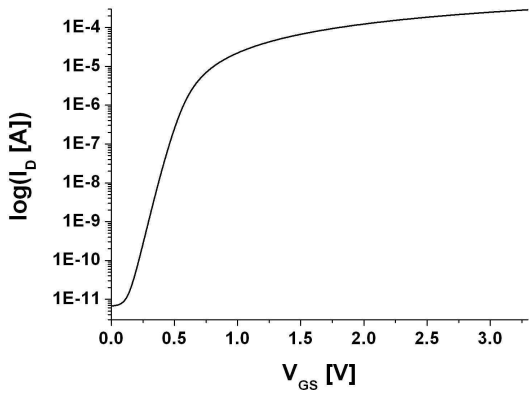

Fig. 3. I_D versus V_{GS} in logarithmic scale – 350 μm

In case of the short-channel device (65 nm), there is the subthreshold exponential region. The current is not dropping to zero immediately for $V_{GS} < V_T$ as it (almost) happens in case of long-channel device (0.35 μm). This is because the distance from source to drain is so short, that in the absence of a conducting channel, the n+ (source) – p (bulk) – n+ (drain) terminals form a parasitic bipolar transistor. As a result, in case of 65 nm technology the static subthreshold current contributes to power dissipation much more than in case of 0.35 μm.

3. Modeling of interconnection line

The chosen model of interconnect lines should meet the needs of simplicity and precision while modeling correctly the distributed character of the circuit. The model is an RC ladder network (Fig. 4) based on the widely accepted π-model [2]. The series inductance is neglected as the corresponding impedance is much smaller than the resistance value for the range of frequencies used in the simulations (from 10 MHz to 0.5 GHz). However as the inductive effect increases, the power dissipation decreases due to the inductance ability to store and release energy [3].

The following parameters of the interconnect line and SPICE models of the transistors for the 65 nm technology have been used [4]:

- $r = 40.7$ kΩ/m,
- $c = 108.3$ pF/m.

Long interconnections are needed for global signals which means they can be found in top layers of a chip. An important assumption in this paper is that there are no coupling capacitances with neighboring lines. The only capacitance is between the path and the lower layers. This assumption makes a significant difference while calculating the value of a capacitance, however all the dependencies are true for both cases.

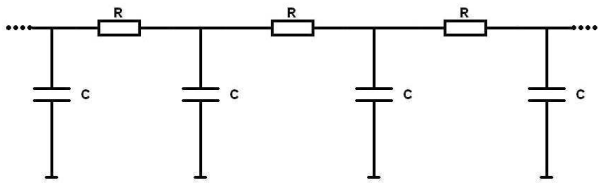

Fig. 4. An RC ladder network

The length of the simulated interconnection is 1 cm. The number of RC ladder rungs was chosen according to time delay modeling calculations. For a 512-rung circuit, the time constant was found with an accuracy of approximately 0.2%. For calculating the temperature distribution the number of rungs was reduced to 64, in order to simplify the calculations. This way the calculated result for the interconnection is much overestimated. As even these overestimated results has negligible values comparing to the values of temperature on the buffers, the simplification does not affect the accuracy.

4. Time delays

Although the length of the global interconnection lines is (relatively) large comparing to the sizes of the elements, the propagation time of it is not the most significant problem. When one considers the time delay of a single inverter and the 1 cm long interconnection line, the delay of the line is much longer (approximately 33 times comparing the small inverter ($W_n = 0.1$ μm, $W_p = 0.26$ μm) with no load and a single line without coupling).

The interconnection driven by an inverter has even longer delay but in this case the crucial problem is the delay on the inverter itself. The buffer which drives the interconnection line has to charge and discharge the parasitic capacitance of the interconnection. The total delay of the circuit becomes far too long to be acceptable. The delay is caused by the interconnection but it effectively takes place on the driving buffer.

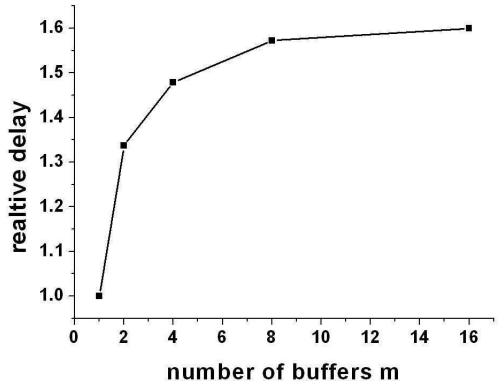

Fig. 5. The influence of the number of buffers on propagation time of the circuit

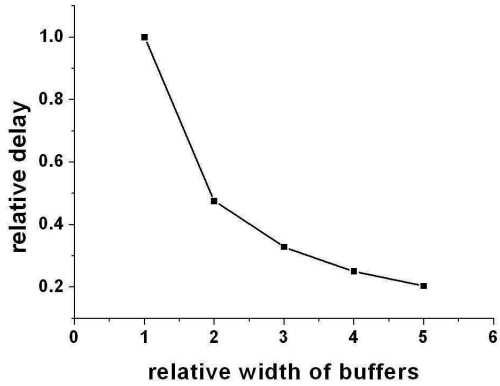

Fig. 6. The influence of the width of buffers on propagation time of the circuit (relative width meaning multiple of $W_n = 0.1$ μm and $W_p = 0.26$ μm)

To make the circuit functional it is crucial to shorten the propagation time of it. The idea is to introduce buffers along the line. They can regenerate the signal and speed up its propagation. The regeneration function of the buffers is very important because the sharpness of the signal's edges influences the speed of switching the inverters. From time point of view the fastest solution is to put only one, very wide buffer at the beginning of the interconnection line. Apparently every additional buffer makes the delay longer (Fig. 5). Nevertheless this disadvantageous dependence weakens with the increasing number of buffers. The more the buffers, the less capacitance each of them has to drive and the sharper the edges of the signal become. The important information is that widening the buffers (changing their aspect ratio) decrease significantly the time delay of the circuit (Fig. 6). As mentioned above, the sharpness of the edges of signals may seriously influence the time delay of the circuit by making the propagation time of the buffers longer.

5. Power dissipation

The aim of the thermal optimization of the circuit is to make the power dissipation as homogenous as possible along the line. However, the simulations show that the power dissipation in the analyzed circuit takes place mostly on the buffer (Fig. 7 and Table 1).

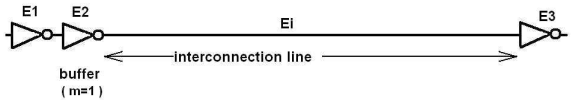

Fig. 7. The energy loss in the circuit – E_1, E_2, E_3 – power dissipations in the buffers, E_i – dissipation in the interconnection

Table 1
The power dissipation in the circuit with a single line

	Energy [fJ]		
	rising edge	falling edge	average
E_1	4.9	2.0	3.5
E_2	530.9	528.1	529.5
E_i	11.0	10.3	10.7
E_3	61.6	61.6	61.6
E_{total}	608.4	602.0	605.2

According to the theory, power dissipation on an inverter is proportional to its load capacitance Eq. (2). Thus the conclusion is that even though the parasitic capacitance of the interconnection line is the main reason of the power dissipation in the circuit, most of this power dissipation takes place on the buffer. The heating of the interconnection's resistance is incomparably smaller and therefore negligible (Table 1). From a thermal point of view, such a heat dissipation distribution is in a complete opposition with what we want to achieve. The major heating is focused in one point, i.e. the location of the buffer.

To change this disadvantageous distribution, one can introduce buffers along the interconnection line (Fig. 8). It has two main advantages: the power dissipation is more homogenously spread in space (the energy losses are more or less equally divided between the buffers) and in time – (e.g. Fig. 9). Introducing buffers does not significantly affect the total energy dissipated in the circuit. The buffers have their own parasitic capacitance but their presence in the circuit reduces the leakage currents and thus even though the total amount of capacitance in the circuit is higher, the energy losses are even lower. The power dissipation is not only more equally spread in the circuit (Fig. 10), but also each buffer dissipates much less energy (Fig. 11).

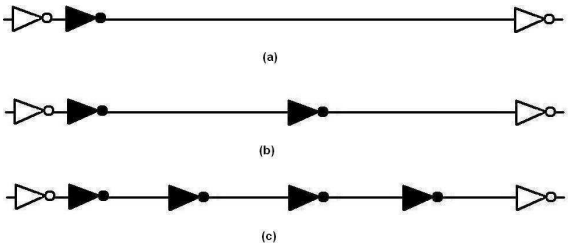

Fig. 8. The number of buffers introduced to the circuit : (a) one buffer $m = 1$, (b) two buffers $m = 2$, (c) four buffers $m = 4$

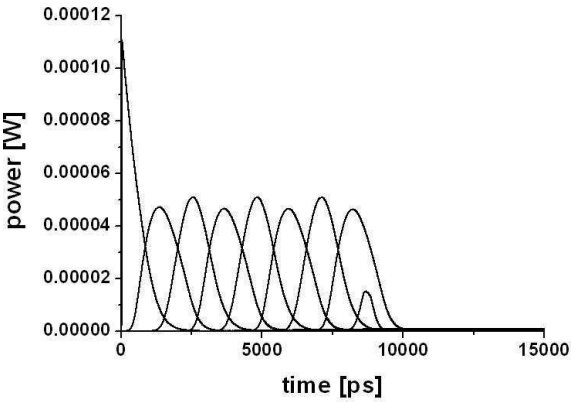

Fig. 9. Power as a function of time of the buffers and the last inverter ($m = 8$ buffers)

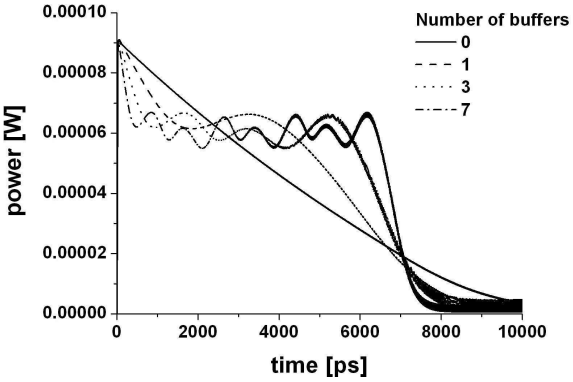

Fig. 10. Total power in the circuit as a function of time and a number of buffers

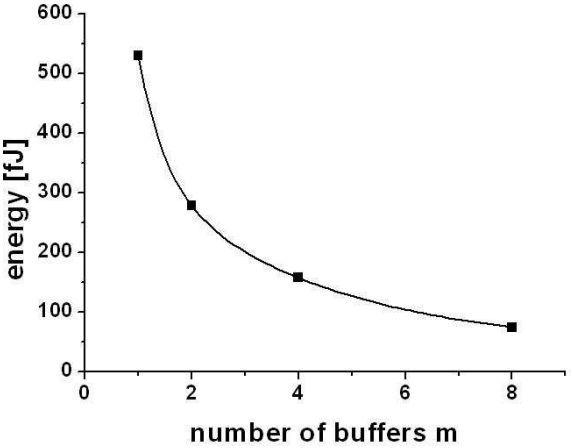

Fig. 11. The average energy on a buffer as a function of number of buffers

Widening the buffers results in lower resistance of the transistors channel. The values of currents are higher and the propagation time is shorter. The total energy losses are not significantly higher, but the power in function of time profile changes (Fig. 12). Due to the higher currents, the power dissipation on the interconnection line is increasing (Fig. 13). The thermal optimization leads us to the conclusion that it is very advantageous to introduce as many buffers as possible. Then the heating would take place along the whole line more homogenously, which is an advantage from thermal point of view. It is essential to notice that these solutions are completely in contradiction with what was recommended for the time delay optimization, where one single buffer was found to be the optimum.

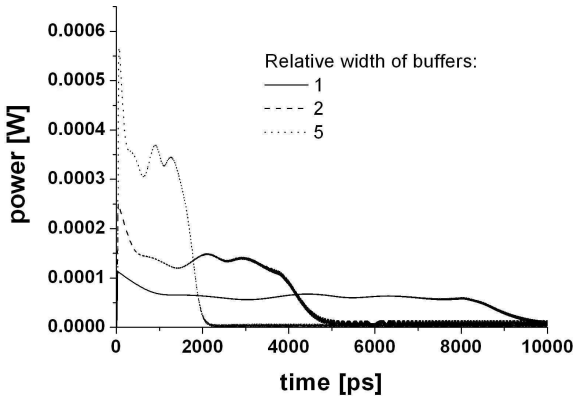

Fig. 12. Total power in the circuit as a function of time and a width of buffers (relative width meaning multiple of $W_n = 0.1\ \mu m$ and $W_p = 0.26\ \mu m$)

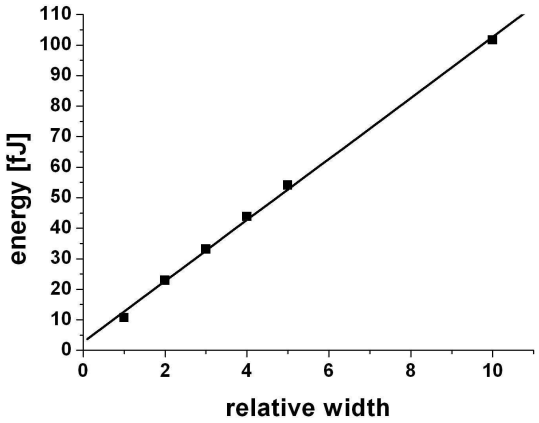

Fig. 13. Energy loss on the interconnection as a function of relative width of buffers (meaning multiple of $W_n = 0.1\ \mu m$ and $W_p = 0.26\ \mu m$)

6. Temperature distribution

The process of distributing heat inside a solid body can be described by the diffusion equation, that in this case is also called the heat equation [5, 6]:

$$k\nabla^2 T = C_v \frac{\partial T}{\partial t}, \tag{3}$$

where k denotes the thermal conductivity and C_v the thermal capacity per unit volume of the solid. For silicon (Si) these parameters are:

- $k = 160$ W/mK,
- $C_v = 1.784 \times 10^6$ J/m^{3}K.

The Eq. (3) has a particular solution known as the Green's function. By definition, the Green's function is the temperature distribution in the solid material if a unit heat pulse occurs at time $t = 0$ in the point $x = y = z = 0$. This particular solution turns out to be:

$$G(r,t) = \frac{\sqrt{C_v}}{8\sqrt{\pi k t}^3} \exp\left(-\frac{r^2}{4\frac{k}{C_v}t}\right), \tag{4}$$

where $r^2 = x^2 + y^2 + z^2$. Remark that (4) turns out to be a three dimensional Gaussian distribution in x, y and z with a rms deviation proportional to $\sqrt{t}$.

In silicon integrated circuits all heat sources are located on the top surface $z = 0$. Only the half space occupied by the solid material can conduct heat, so that (Eq. 4) has to be doubled to get the correct temperature distribution: $G \to 2G$. As outlined in the previous sections, the heat production is distributed spatially (in transistors and interconnection lines) but varies in time as well. Generally, the heat production on the surface $z = 0$ p expressed in W/m^2 is a function of x, y and t: $p(x, y, t)$. The resulting temperature distribution is then:

$$T(x,y,z,t) = \\ = \int_0^t dt' \iint p(x',y',t')\ 2G(r,t-t')\ dx'dy', \tag{5}$$

where $r^2 = (x-x')^2 + (y-y')^2 + z^2$. In our calculations, we are mainly interested in the temperature distribution on the top surface so that $z = 0$ can be substituted in Eq. (4).

An important feature of a material is its thermal diffusivity α [7]:

$$\alpha = \frac{\text{heat conducted}}{\text{heat stored}} = \frac{\text{k}}{\text{C}_v}\left[\frac{\text{m}^2}{\text{s}}\right]. \tag{6}$$

A physical interpretation of α is that it tells whether a material has the ability of conducting or storing heat. As a result it is a measure of how fast heat diffuses through a material. It appears that silicon (Si) is quite a good heat conductor and tends to conduct the heat rather than to store it (Table 2).

Table 2
Values of thermal diffusivity for different materials

material	α [m^2/s]
silicon	90×10^{-6}
water	0.14×10^{-6}
silver	174×10^{-6}

A circuit with 4 buffers (Fig. 14) has been analysed if a input square wave of 1 GHz is applied. The power dissipation is shown in Fig. 14. As expected the pattern is repeated periodically every 0.5 ns. The resulting temperature in each of the buffers has been evaluated using (Eq. 4) and also plotted in Fig. 14. At first sight the temperature varies just like the power dissipation. A closer look however, reveals that the temperature shows a small delay as compared to the power dissipation. It is also remarkable that temperature peaks up to 30 degrees above ambient are observed.

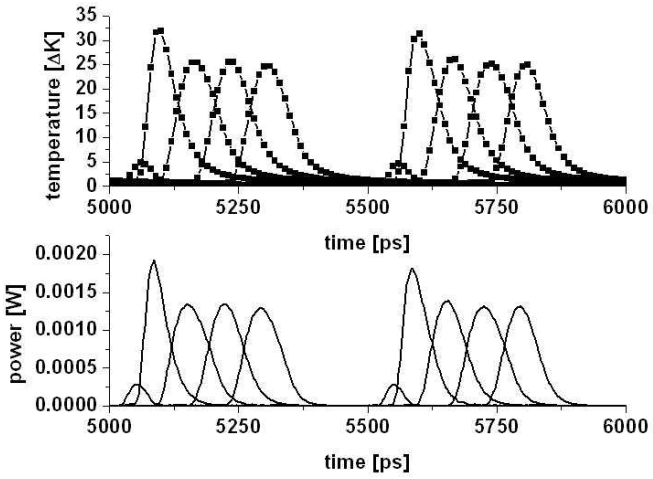

Fig. 14. Power dissipation and temperature on inverters, circuit with four buffers, width ($W_n = 3$ μm and $W_p = 8.14$ μm), $f = 1$ GHz

Another view on the same results is shown in (Fig. 15) where the temperature of the first buffer has been displayed versus time but with a different temperature scale. The peak of 30 degrees are then no longer visible but one clearly sees now the slow increase of the offset temperature. This is due to the fact that Eq. (6) assumes that $T = 0$ in the beginning. This offset temperature would also occur if the first buffer would have a constant power dissipation from $t = 0$ on, provided that the total energy remains the same. The fact that this offset temperature (0.5 degrees) turns out to be much smaller than the peak values, proves that a time dependent thermal analysis of integrated circuits should be carried out carefully. Neglecting fast power transients is no longer permitted if one wants to know the peak temperatures in nanoscale transistors.

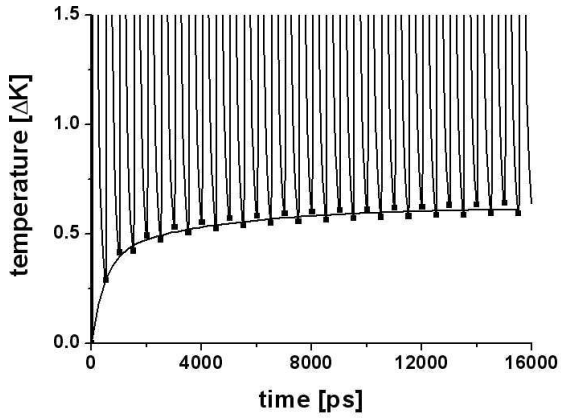

Fig. 15. Temperature that settles on the first buffer, circuit with four buffers, circuit with four buffers, width ($W_n = 3$ μm and $W_p = 8.14$ μm), $f = 1$ GHz

It is necessary to point out that the phenomena related to vias are not taken into account though they may affect the temperature distribution depending on the via density [8].

7. Conclusions

Although the sizes of single CMOS components continously decrease, the sizes of chips remain unchanged. The density of the circuit increase and the length of the global interconnection has to remain unchanged. To drive properly the interconnection it is important to focus on the question of frequency and synchronization of signals. It is also vital to analyze and take into account thermal aspects of the circuit.

Introducing buffers along the line can help to reduce the time delay and to make the power dissipation more homogenous. Consequently the temperature rises decrease. It is due to the fact that placing buffers at equal intervals makes each of them to drive a smaller (divided by the number of buffers) capacitance. Thus it is advisable to introduce a number of buffers along the line. They reduce the time delay, ensure the homogeneity of power dissipation and therefore influence the temperature distribution.

The peak values of the power transients are much larger than the offset temperature which proves that a time dependent thermal analysis of nanoscale integrated circuits should be carried out carefully.

Acknowledgements. K. Gnidzinska wants to thank the EU for the financial support within the framework of the Erasmus Socrates student exchange program.

REFERENCES

[1] *International Technology Roadmap for Semiconductors*, http://www.itrs.net/, 2005.

[2] A. Chatzigeorgiou, S. Nikolaidis, and I. Tsoukalas, "Modeling CMOS gates driving RC interconnect loads", *IEEE Trans. Circuits and Systems II: Analog and Digital Signal Processing* 48 (4), 413–418 (2001).

[3] Y.I. Ismail, E.G. Friedman, and J.L. Neves, "Dynamic and short-circuit power of CMOS gates driving lossless transmission lines", *IEEE Trans. Circuits and Systems I: Analog and Digital Signal Processing* 46 (8), 950–961 (1999).

[4] *Predictive Technology Model (PTM)*, http://ptm.asu.edu/, 2007.

[5] B. Vermeersch and G. De Mey, "Thermal impedance plots of micro-scaled devices", *Microelectronics and Reliability* 46 (1), 174–177 (2006).

[6] G. De Mey, *Various Applications of the Boundary Element Method*, Editura Universităţii din Oradea, Oradea, 2002.

[7] Y.A. Çengel, *Heat Transfer: a Practical Approach*, McGraw-Hill, London, 2002.

[8] T.-Y. Chiang, K. Banerjee, and K.C. Saraswat, "Analytical thermal model for multilevel VLSI interconnects incorporating via effect", *IEEE Electron Device Letters* 23 (1), 31–33 (2002).