End of Moore's law: thermal (noise) death of integration in micro and nano electronics

Laszlo B. Kish [1]

Texas A&M University, Department of Electrical Engineering, College Station, TX 77843-3128, USA

Abstract

The exponential growth of memory size and clock frequency in computers has a great impact on everyday life. The growth is empirically described by Moore's law of miniaturization. Physical limitations of this growth would have a serious impact on technology and economy. A thermodynamical effect, the increasing thermal noise voltage (Johnson–Nyquist noise) on decreasing characteristic capacitances, together with the constrain of using lower supply voltages to keep power dissipation manageable on the contrary of increasing clock frequency, has the potential to break abruptly Moore's law within 6–8 years, or earlier.
© 2002 Elsevier Science B.V. All rights reserved.

1. Introduction

In 1965 Moore observed that integrated circuit complexity evolved exponentially [1]. As a consequence of this observation a scaling algorithm was developed in the 1970s, stating that device feature sizes would decrease by a factor of 0.7 every three years. Though this empirical law has attracted various kinds of critics (see, e.g., [2,3]), this prediction has proven to be accurate enough that it has become a solid-state electronics industry target that must be met by semiconductor device manufacturers in order to remain competitive. The silicon industry has been following Moore's law, and transistor sizes on chips have been shrinking exponentially [4]. Today characteristic sizes are around 100 nanometers, and the trend is believed to continue for the far future [5–7]. News claim transistor sizes down at 9 nm by leading research labs [6], and there are new integration technologies with a potential to follow Moore's law for, at least, another two decades [7]. Moore himself believes in some decrease of the exponent of growth due to economical reason [8].

On the contrary of the above mentioned results and expectations, in this Letter, we shall point out that further increase of the integration density of computer chips may face a physical limit, abrupt and major complications due to false bit occurrences generated by thermal fluctuations (Johnson–Nyquist noise). The expected range of sizes where serious problems will emerge is around 40 nanometers and below, which means serious miniaturization problems may be expected in 6–10 years.

E-mail address: laszlo.kish@ee.tamu.edu (L.B. Kish).
[1] Previously L.B. Kiss.

The effect causing the problem is due to power dissipation and the energy equipartition theorem in thermodynamical systems, so it is fundamental and general. It does not set any physical limit on possible transistor sizes but on the integration density on a chip of fixed maximal dissipation. The only way to get around this effect would be either to give up increasing the integration density, that is itself Moore's law, or to give up increasing the clock frequency. However, increasing the clock frequency of one of the most important reasons for miniaturization.

2. Speed, size and dissipation

In this section, we survey some well-established practical facts about speed, dissipation, capacitance, and size. At the following considerations, we focus on digital (C)MOS technology because it is the most widely used one. However, the same considerations can be used for other technologies with very similar outcome. During the considerations, we assume that we run the devices with the highest possible clock frequency they are fast enough to handle.

(C)MOS technology has the advantage that its power dissipation is the smallest among all. The reason for that is simple. When the devices are in a certain logical state, the current consumption is negligible. So, in the first approach, all dissipation takes place during transitions between logical states and the source of this dissipation is the need to charge or discharge the related capacitors. On Fig. 1, a simple circuit model of the power dissipation related to the gate capacitance of a single transistor is shown. C is the capacitance of the gate, R is the drain-source resistance of the other (driving) transistor(s) which changes the logic state of the gate. U_0 is the supply voltage of the circuit.

From this picture, some well-known relations follow.

The upper limit of the clock frequency f_0, where the alternating voltage driving is still sufficient to switch the transistors between logical states, can be given as

$$f_0 \cong (RC)^{-1}. \tag{1}$$

The energy dissipated by this unit during one period of the clock frequency is scaling as CU_0^2. The average
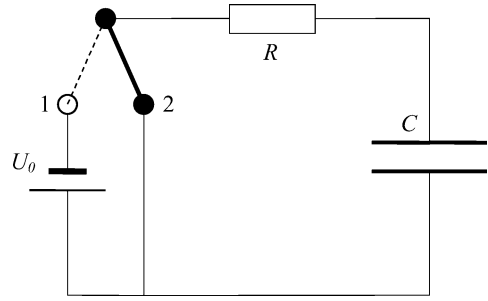

Fig. 1. *RC* model to estimate the speed and power dissipation in digital MOS integrated circuits. U_0 is the supply voltage and R is the characteristic drain-source channel resistance. C is the gate capacitance. The switch is alternating between the supply voltage and the ground by the clock frequency.

power dissipation of this single unit (Fig. 1) is scaling as

$$P_1 \propto f_0 E_1 \propto (RC)^{-1} CU_0^2 \propto \frac{U_0^2}{R}. \tag{2}$$

Note, the clock frequency does not appear in this formula due to the assumption that we run the unit at the highest clock frequency f_0.

However, the total power dissipation P_N of a IC chip with fixed geometrical size containing N such units depends on N, R and C. To show the impact of miniaturization on the power dissipation and clock frequency of an IC chip with fixed size, first we have to determine the scaling of N, R and C versus the characteristic transistor size s. Obviously, N is scaling as $1/s^2$. Using the reasonable assumption that the depth of drain-source channel is constant, R can be approximated as constant. From these considerations, we obtain that P_N is scaling as

$$P_N \propto \frac{NU_0^2}{R} \propto NU_0^2 \propto \frac{U_0^2}{s^2}. \tag{3}$$

Relation (3) has an exact match with data describing the past and expected future of the evolution of microelectronic technology [4]. Therefore, if we want to keep the dissipation constant during miniaturization and also the dense packing of transistors, Eq. (3) requires the supply voltage to scale linearly with the size.

To calculate the thermal noise, we will need to know the capacitance. If the thickness of the gate oxide is constant then C is scaling as

$$C \propto s^2. \tag{4}$$

If, instead of constant gate thickness, we suppose that the gate oxide thickness is linearly scaling with s, we get

$$C \propto s. \tag{5}$$

The practical situation has been between the two limits given by Eqs. (4) and (5) because the gate oxide thickness has gradually been shrinking, however not so fast as the characteristic size of the transistors on the chip [4].

3. Thermal noise voltage in the whole frequency band

Thermal noise in resistors is a stationary Gaussian stochastic voltage fluctuation process with zero mean value. The well-known Johnson–Nyquist formula describes the power density spectrum S_u of the thermal noise voltage in the low frequency limit

$$S_u(f) = 4kTR, \tag{6}$$

where f is the frequency, k the Boltzmann constant, T the temperature and R the resistance. If the resistance is shunted by a capacitor then Eq. (6) and linear network theoretical calculations yield that the total root-mean-square (RMS) thermal noise voltage U_n on the capacitor is

$$U_n = \sqrt{\frac{kT}{C}}. \tag{7}$$

Eq. (7) can easily be obtained also from the energy equipartition theorem of equilibrium thermodynamics. As the capacitor has one thermodynamical degree of freedom, the mean energy in it should be $kT/2$ what requirement directly implies Eq. (7). This is the physical reason why the value of R does not appear in Eq. (7).

Eqs. (4), (5) and (7) imply that, when the full bandwidth is used like in digital circuits, the RMS thermal voltage is steadily increasing during miniaturization. This effect has not been widely realized yet, probably because of Eq. (6) what indicates a constant power density spectrum when the resistance is kept constant. The increasing bandwidth, $(RC)^{-1}$, implies the increase of the RMS noise voltage even if its low-frequency spectrum remains unchanged. This increase evolves between the scaling limits of $1/s$ and $1/s^{0.5}$.

4. False bit-flips due to thermal noise

The mean frequency $\nu(U_{\text{th}})$ of crossing a threshold amplitude limit U_{th} by a Gaussian noise process of zero effective value is given by the Rice formula [9–11]

$$\nu(U_{\text{th}}) = \frac{2}{U_n} \exp\left(\frac{-U_{\text{th}}^2}{2U_n^2}\right) \sqrt{\int_0^\infty f^2 S(f)\, df},$$

$$\text{where} \quad U_n = \sqrt{\int_0^\infty S(f)\, df}. \tag{8}$$

For band-limited white noise, $S(f) = S(0)$ for $f \leqslant f_c$ and $S(f) = 0$ for $f_c < f$, the formula becomes

$$\nu(U_{\text{th}}) = \frac{2}{\sqrt{3}} \exp\left(\frac{-U_{\text{th}}^2}{2U_n^2}\right) f_c,$$

$$\text{where} \quad U_n = \sqrt{S(0) f_c}. \tag{9}$$

Whenever the thermal noise causes crossing of the value of the actual logic threshold voltage, which is the noise margin between the actual voltage difference between the two logic values, a false bit flip occurs. It is obvious from Eq. (9) that, in the case of low noise $(U_{\text{th}}/U_n \gg 1)$, the explicit U_{th}/U_n ratio dominates the frequency of false bit-flips. This fact can also be seen on Fig. 2 where the frequency of false bit occurrences is plotted versus U_{th}/U_n. The clock frequency is assumed to be equal to f_c. Apparently, the practical difference between the $f_c = 2$ GHz and $f_c = 20$ GHz cases is negligible, especially, when we compare this dependence to the steep variation versus U_{th}/U_n. Similarly, the number of transistors in the computer has also a minor role at determining the necessary noise margin. It is obvious from Fig. 2 that U_t has to be greater by U_n about an order of magnitude for the reliability of the computer. Today's PCs have about 5×10^9 transistors; 10^8 are in the processor and the rest in the RAM and the other ICs. In a decade, the clock frequency will approach 20 GHz and the number of transistors will grow further order of magnitude.

So in conclusion, it is reasonable to say that a safe thermal noise margin requires

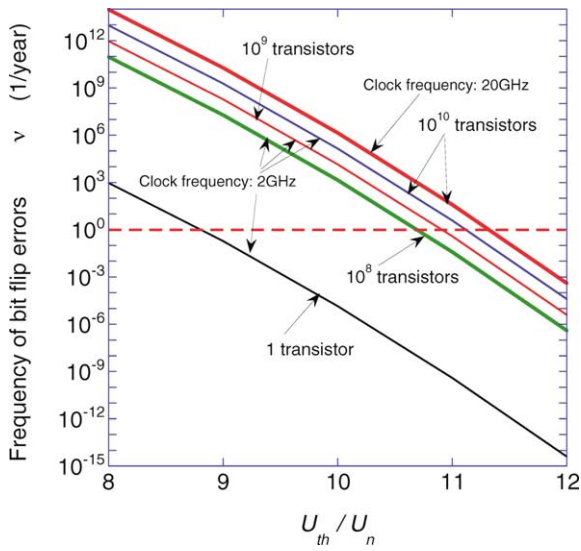$$U_{\text{th}}/U_n \geqslant 12 \tag{10}$$

in modern digital circuits.

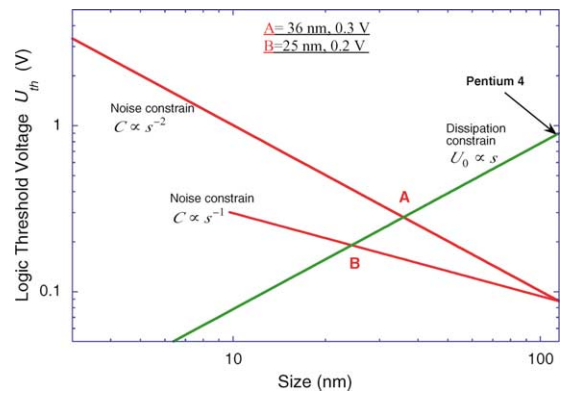Fig. 2. Frequency of false bit-flips due to thermal noise.

Fig. 3. Evolution of the logic threshold voltage U_{th} due to constrains of dissipation and noise. Noise constrain is a lower limit, dissipation constrain is an upper limit. Moore's law dies when the miniaturization gets between A and B. Due to the lack of information, this calculation is based on the assumption $U_{th} = 0.6\,U_0$, which is very optimistic, see in the conclusion.

At the above considerations, error-correcting codes were not taken into account. One reason for that is the sake of simplicity, another reason is the trend toward more cache processing in complex logic integrated circuits, which reduces the advantages of error correcting codes [5].

5. Limitations of Moore's law in complex digital integrated circuits

In this section, we apply the results obtained in Sections 2–4 for estimation of the practical situation with thermal noise. The logical threshold voltage $U_{th} = U_{high\,min} - U_{low\,max}$, where $U_{high\,min}$ is the minimal legal voltage value of the logic high state and $U_{low\,max}$ is the maximal legal voltage value of the logic low state. Following the old 5 V supply voltage practice, we assume that scales linearly with the supply voltage, $U_{th} = 0.6\,U_0$. Note, due to offset and crosstalk problems, the fraction 0.6 may be too optimistic for future practice, however, we have been unable to find any prediction how this fraction would diminish during miniaturization, so we stick to this optimistic value. The main results of this paper are shown on Fig. 3.

The dissipation constrain curve is based on relation (3) and it shows how the U_{th} should be decreased

versus the decreasing s in order to avoid increasing dissipation. The curve indicates the *allowed maximal* value of U_{th}. The right end of the curve represents today's data [4] (2002), $s = 115$ nm and $U_{th} = 0.6 \times 1.5$ V $= 0.9$ V.

The noise constrain curves are based on relations (4), (5), (7) and (10) and they show how the U_{th} should be increased versus the decreasing s in order to avoid false bit-flips. The curves indicate the *required minimal* value of U_{th}, in the cases when relations (3) or (4) is valid, respectively. The real condition is between these two limits. The right end of the curves represent today's data [4] (2002), physical gate length of 53 nm, gate width/length aspect ratio of 2.5, gate thickness of 2 nm. A relative permittivity value of 3 was assumed for the gate oxide.

The dissipation constrain curve intersects the noise constrain curves at points A (36 nm) and B (25 nm). Somewhere between these points, either the noise constrain or the dissipation constrain cannot be satisfied, because the requirements are contradictory. Taking Moore's law into account, a rough estimate shows that we will reach the danger zone in 6–8 years.

6. Conclusions

In this Letter, we have studied potential technological difficulties of future miniaturization with high-

density integration due to a fundamental an unavoidable noise process, the thermal noise. Our aim has been to take the attention to the problem of thermal noise, which has the potential of breaking Moore's law.

It is important to emphasize here the nature of the problem. The arguments in this Letter do not set any physical limit on transistor sizes. What is claimed here, the logical threshold voltage, therefore the supply voltage, cannot be reduced below a certain limit due to false bit-flips induced by thermal noise. The only way to get around this effect would be either giving up to increase the integration density, that is itself Moore's law, or giving up to increase the clock frequency. However, that would be contradictory because increasing the clock frequency is one of the most important reasons for the miniaturization efforts.

For the sake of simplicity and generality, we used several strong approximations during these considerations. Most of the approximations were optimistic, so the real situation is probably worse than Fig. 3 shows. We used the following, optimistic simplifications.

(1) Noises due to switching of devices (ground plane EMI), which are major factors in logic IC design, have been neglected.

(2) Excess thermal noise due to hot electrons has not been taken into consideration.

(3) The logical threshold voltage was supposed to be 60% of the supply voltage, which may be too high. Due to offset problems with the gate threshold voltage, this ratio will probably diminish while the supply voltage is decreased.

(4) When applying the Rice formula for threshold crossing, we supposed that the clock frequency is the relevant bandwidth. However, in the nominator in the Rice formula, what contains the RMS velocity, the relevant bandwidth is always greater, it is determined by the second pole of the transfer function. So, the actual probability of threshold crossing is greater than supposed here.

(5) Other noise phenomena (shot noise, $1/f$ noise) were neglected.

Some of the approximations were pessimistic.

(i) Parasitic capacitors were neglected. These capacitors can increase gate capacitance by 20–30% [4].

(ii) Error-correcting codes were not taken into consideration. As we mentioned above, there is a trend toward more cache processing in complex logic integrated circuits, which reduces the advantages of error-correcting codes [5]. Furthermore, note that a 10% increase of the U_n/U_{th} ratio causes a factor of -1000 increase in the false bit frequency. The very progressive growth of false bit-flip frequency at a small increase of the thermal noise or a small decrease of threshold voltage indicates that error-correcting codes cannot save Moore's law when the danger zone is reached.

Even the simplified considerations make it clear that the effect of thermal noise can become a serious issue of high-density integration at the beginning of the next decade, and the danger zone is approached already within this decade. The potential impact on future nanoelectronics technology is also evident. It is important to note that, due to the optimistic sides of our estimation, there is a possibility that the described effects are already causing problems in the research and development of highly integrated logical ICs. As the false bit-flips occur at random locations and times, and the number of possible locations is great, it is not easy to identify the source of their origin.

Acknowledgements

References

[1] G.E. Moore, Electronics 38 (8) (1965).
[2] I. Fodor, Phys. Today 53 (2000) 106.
[3] A.E. Brenner, Phys. Today 54 (2001) 84.
[4] International Technology Roadmap for Semiconductors, Semiconductor Industry Association, 2001, http://public.itrs.net.

[5] L. Condra, The Impact of Semiconductor Device Trends on Aerospace Systems, Report unpublished.

[6] Global Sources, July 11, 2002, http://www.globalsources.com/MAGAZINE/EC/0209W2/TUTSMC.HTM.

[7] B. Fuller, Moore's law scales to at least 9 nm, EE Times, March 12, 2002, http://www.eetimes.com/semi/news/OEG20020312S0077.

[8] L. Neely, Electronic News, July 10, 2002, http://email.electronicnews.com/cgi-bin2/flo?y=eHha0EjxX60DbD0-Boi10Ag.

[9] S.O. Rice, Bell Syst. Tech. J. 23 (1944) 282;
S.O. Rice, Bell Syst. Tech. J. 24 (1945) 46.

[10] L.B. Kiss, in: R. Katz (Ed.), Chaotic, Fractal, and Nonlinear Signal Processing, American Institute of Physics, 1996, p. 382.

[11] B. Kedem, Time Series Analysis by Higher Order Crossings, IEEE, New York, 1994.