Temperature Scaling of CMOS Circuit Power Consumption

Victor Sverdlov [a,1], Yehuda Naveh [b], Konstantin Likharev [a]

[a] *Department of Physics and Astronomy, SUNY Stony Brook, Stony Brook, NY 11794-3800, USA*

[b] *IBM Research, Haifa 31905, Israel*

Abstract

We have analyzed fundamental physical limitations on power consumption of prospective semiconductor digital integrated circuits based on nanoscale silicon MOSFETs, using simple models of these devices and power dissipation. Results show that the temperature dependence of the power is determined by circuit speed requirements. For high speed operation, both power P and power supply voltage V_{DD} saturate when T is reduced below approximately 100 K. In the low speed limit, P scales as T^2, while V_{DD} drops linearly with T. However, thermal fluctuations may alter this scaling, leading to $P \propto T$ and $V_{DD} \propto T^{1/2}$, at low temperatures and/or large circuit densities. We compare this scaling with that of superconductor RSFQ logic.

Key words: MOSFET; RSFQ; CMOS circuits; electron devices

1. Introduction

Continuing aggressive scaling of MOSFET critical dimensions into the sub-100-nm range and the resulting circuit density growth lead inevitably to a further increase of power consumption of CMOS integrated circuits. This increase presents one of the main challenges for the further successful progress of this leading electronic technology [1–3]. One way to reduce the power consumption could be circuit cooling. However, as temperature is reduced, the maximum heat density removed from a chip by comparable technical means is also decreased, at least proportionally to T. This is why in most cases cooling is effective only if power dissipation decreases faster than T.

In order to estimate the temperature dependence of the minimum power, we have combined a 1D model [4,5] of the most promising transistor type, a double-gate, nanoscale MOSFET with an ultrathin intrinsic channel, with a simple model for total (static plus dynamic) power consumption in CMOS integrated circuits. In the later model [6], the specific power P (per unit channel width) may be expressed as a function of just three parameters: the power supply voltage V_{DD}, linear current density J_{ON} of an open transistor, and circuit "activity parameter" $\lambda \equiv 2\pi p \sum_i \alpha_i W_i / W$. Here p is the ratio of the load recharging time constant to the clock period, α_i is the switching activity of the i-th circuit block, W_i is the width of all logic transistors in the i-th block, and W is the width of all transistors of the circuit. For any particular transistor model, the function $P(J_{ON}, V_{DD}, \lambda)$ can be minimized over V_{DD}, yielding the minimum power and optimum power supply voltage (Fig. 1).

2. Results

Calculations show that temperature behavior of the minimum power depends strongly on circuit speed assumptions. In high-performance logic circuits, current density J_{ON} has to be quite high (close to 600 A/m, the industrial standard). In this case, both the minimum power and the optimum V_{DD} reduce with temperature rather slowly and saturate close to 100 K (see

[1] Corresponding author. Present address: Department of Physics and Astronomy, SUNY Stony Brook, Stony Brook, NY 11794-3800, USA. E-mail: svk@hawk.physics.sunysb.edu

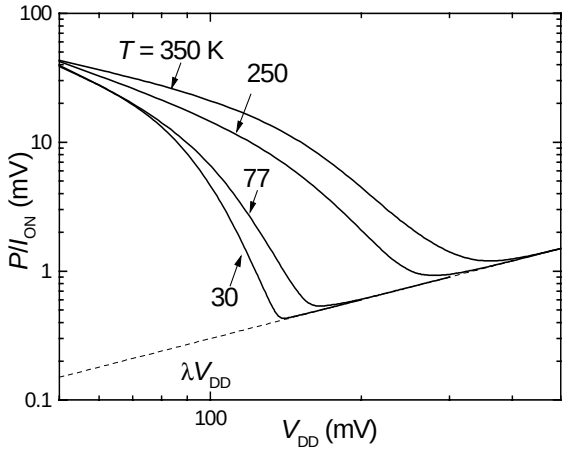

Fig. 1. Total power of a CMOS circuit as a function of V_{DD} for several values of T, at fixed J_{on} =600 μA/μm. Transistor parameters: electrode doping $3 \times 10^{20} cm^{-3}$, channel thickness 3 nm, gate oxide thickness 1.5 nm.

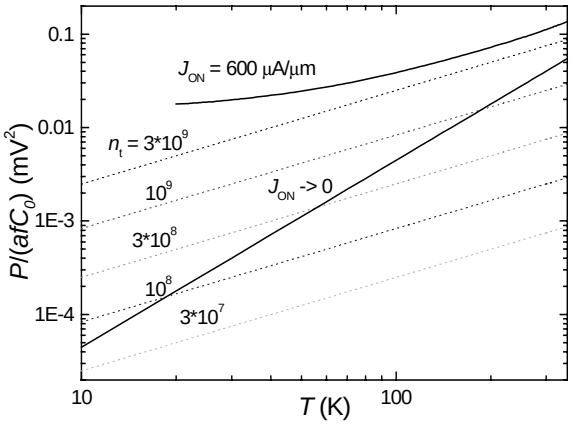

Fig. 2. Power P as a function of temperature. Solid lines: minimum power for high-performance (upper curve) and low-speed operation. λ= 0.003. Dashed lines are lower bonds for P set by thermal fluctuations, for several n_t (in cm^{-2}) and $C_0 = 10^{-7}$F/cm$^-$2. Transistor parameters are the same as in Fig. 1.

the upper solid line in Fig. 2). This saturation is cased by the need in a high gate voltage swing V_{DD} necessary to bring the transistor from a well closed state to a well open state. This voltage is determined by the density of 2D electron states in the channel, and its specific capacitance [4,5] and for high source and drain doping does not depend on temperature.

If the high speed requirement is removed, and J_{ON} may be small, the lowest power is achieved in the subthreshold region. In this case, the ratio $P/(J_{ON}T^2)$ is a function of the activity parameter λ alone, so that the minimum power P scales as T^2 (strait solid line in Fig. 2). In this case, the optimum power supply voltage V_{DD} is proportional to T.

However, this scaling may be invalid because of the thermal fluctuations. Indeed, the energy scale of a CMOS signal, $E = C_0 V_{DD}^2/n_t$, (where C_0 is the total load capacitance per unit area, and n_t is transistor density) must be much larger than the scale of thermal fluctuations, $k_B T$. (In order to keep bit error rate to acceptable level, the ratio $E/k_B T$ should not be below 300.) Dashed lines at the Fig. 2 show the limit for the minimum power set by thermal fluctuations, for several values of n_t. This figure shows that for the circuit densities above 10^9 transistors per cm^2, envisioned for the end of this decade [1], the fluctuation-imposed power limit is not too much below than that for high-speed operation.

Thus, deep cooling of CMOS (and more generally, all semiconductor transistor) integrated circuits may be advantageous only in special cases of relatively low speed and low density circuits. In contrast, low-temperature superconductor RSFQ circuits (see, e.g., the recent review [7]) with submicron junctions may combine very high operation speed (beyond 100 GHz) with extremely low power, of the order of 100 nW/gate. The fundamental reason for this difference is that RSFQ devices, which communicate with ultrashort pulses propagating ballistically along superconductor transmission lines, need not change the physical state of the interconnect as a whole each clock cycle, while in semiconductor circuits with their normal-metal wiring such change is apparently unavoidable.

Acknowledgements

The work was supported in part by the Semiconductor Research Corporation.

References

[1] International Technology Roadmap for Semiconductors. 2001 Edition. Available on the Web at public.itrs.net/Files/2001ITRS/Home.htm.

[2] D.J. Frank, R.H. Dennard, E. Nowak, P.M. Solomon, Y. Taur, and H.-S.P. Wong, Proc. IEEE **89** (2001) 259.

[3] P.P. Gelsinger, In: 2001 ISSCC Tech. Dig. (IEEE Press, Piscataway, NJ, 2001) 22.

[4] K. Natori, J. Appl. Phys. **76** (1994) 4879.

[5] K. Likharev, "Sub-20-nm Electron Devices", to be published.

[6] V. Sverdlov, Y. Naveh, and K. Likharev, In: 2001 ISDRC Proceedings (IEEE Press, Piscataway 2001) 547.

[7] P. Bunyk, D. Zinoviev, and K Likharev, Int. J. of High Speed Electron. and Syst **11** (2001) 257.