Extended Abstracts of the 1996 International Conference on Solid State Devices and Materials, Yokohama, 1996, pp. 1-3

Invited

Future CMOS Scaling-Approaching the Limits?

Robert H. Dennard
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 USA

Future trends in CMOS technology are examined in the context of a generalized scaling approach. Problems in scaling threshold voltage and wiring dimensions are beginning to limit the traditional benefits.

1. INTRODUCTION

Scaling of microelectronic devices and circuits to smaller and smaller dimensions has been amazingly successful since we introduced some of the principles in the early 1970's [1]. The deep submicron CMOS of today is projected to be scaled down to less than 100 nm channel lengths in the next ten years. However, we are approaching some of the fundamental limits of scaling.

Our present concept of scaling, shown in Table I, has been broadened from the original where the electric field was kept constant and the devices and wires were scaled together. Most device physical dimensions are divided by a factor of α_d, while the electric field is allowed to be multiplied by a factor ε so that the voltage can be reduced more gradually than the device dimensions [2]. The wiring dimensions and the device width are divided by a factor α_w. Even if the field ε increases, a reasonable goal is to increase the circuit speed by a factor α_d (which assumes the average carrier velocity remains about the same). At that speed, the active power for a given circuit scales as $\varepsilon^2/\alpha_d\alpha_w$ while the power density scales as $\varepsilon^2\alpha_w/\alpha_d$.

Physical Parameter	Generalized Scaling Factor
Channel Length, L	$1/\alpha_d$
Gate Insulator, t_{ox}	$1/\alpha_d$
Voltage, V	ε/α_d
Wiring Width	$1/\alpha_w$
Channel Width, W	$1/\alpha_w$
Circuit Speed (goal)	α_d
Circuit Power	$\varepsilon^2/\alpha_d\alpha_w$

Table 1. Generalized scaling approach.

These results assume that the device threshold voltage can be scaled along with the power-supply voltage. However, it is well known that scaling down the threshold voltage (V_t) tends to increase leakage current (I_l) in the turned-off devices. Even if V_t is not scaled, the I_l current per device goes up approximately as C_{ox}W/L [3]. In the context of scaling, this means I_l per circuit would scale as α_d^2/α_w. Multiplied by the number of circuits per unit area, α_w^2, this gives a standby current per unit area of $\alpha_d^2\alpha_w$. Fortunately, the standby current is quite low in today's 3.3V chips. The worst values occur for the shortest devices at the highest operating temperature, while the worst performance is determined by the longest devices as illustrated in Fig. 1. Thus by tightening tolerances, the threshold V_{t+} can be reduced some in future scaled devices, while the worst-case threshold V_{t-} may be reduced less if at all. Nevertheless, leakage currents are expected to increase (as described above) for constant V_{t-} and will increase further by about 10X for each 100 mV lowering of V_{t-}. This is a key limit to future CMOS scaling.

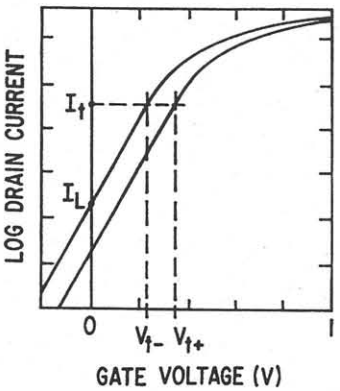

Fig.1. Illustration of tolerance limits of V_t.

Many schemes have been discussed which use variable body voltage to adjust V_t so as to minimize leakage in worst-case tolerance conditions or for test or standby modes. These have a cost in complexity and density, and the resulting improvement appears to be limited.

2. PERFORMANCE/POWER PROJECTIONS

A proposed scaling path for CMOS technology for applications such as microprocessors has been previously described [4]. Two scenarios have been defined, one optimized for high performance (HP) and another for much lower power dissipation (LP).

Figure 2 shows delay of a typical loaded NAND circuit versus channel length for these two scenarios. Both cases have the oxide thickness scaling down along with the channel length and have been demonstrated with experimental hardware. The HP case has the power-supply voltage gradually scaled down such that the electric field factor, ε, increases substantially as L and t_{ox} are decreased. For the LP scenario, the voltage is scaled down more rapidly so the electric field only rises slightly more than in the original "constant electric field" scaling formulation. For the HP case, V_t is scaled down somewhat less than the applied voltage in order to keep the standby current reasonable. Low standby current is important to allow I_{DDQ} testing to be carried out at room temperature, to allow reasonably low d-c power consumption at worst-case operating temperature, and to allow functional operation at

more elevated burn-in temperature and voltage. The same V_t values are used for the LP case so standby power can be kept reasonably low for battery-powered applications.

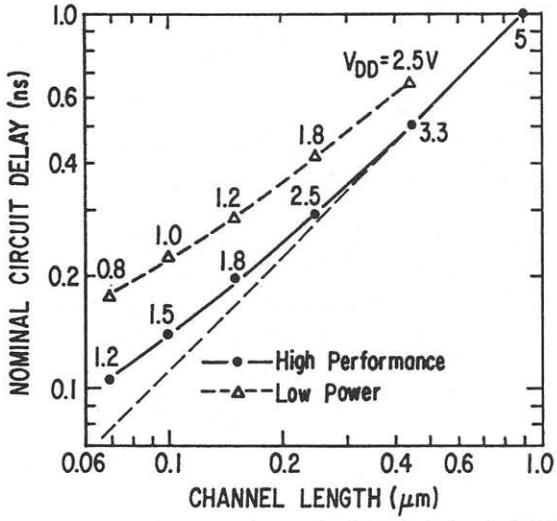

Fig. 2. Delay projections for typical CMOS loaded NAND.

The dashed line in Fig. 2 shows the goal of having the circuit delay reduce in proportion to channel length. The high-performance case departs gradually from this goal, while the low-power case loses more performance and is about 2X slower at the smallest channel lengths. This is directly related to the non-scaling of the threshold voltage in both cases, along with other small factors in the low-power (low-voltage) case.

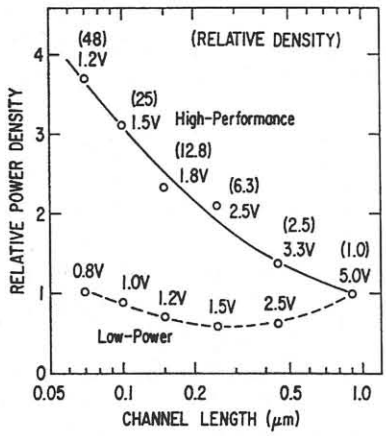

Fig. 3. Power density of scaled CMOS.

The power dissipation is illustrated in Fig. 3. The power is closely related to the wiring density which is noted. The value of 6.3 quoted for the 2.5V HP technology represents today's planarized 5-level interconnection technology for 0.5μm lithography with 1.8μm pitch for middle levels M_2-M_4. Future density increases are projected as 2X improvement per device generation. For the HP scenario, the capacitance increase due to denser wire and the speed increase drive up the power density since the voltage is scaled only gradually. The LP design curve has much better power density due to the lower voltages and the lower switching frequency.

3. SCALED INTERCONNECTION LINES

The original paper on scaling identified two problems with scaled interconnection lines. One is the RC delay which remains constant when lines are scaled in all dimensions (including length) and may impose a limitation on speed. The other problem is the increased current density due to the smaller cross section. It is interesting to consider whether these pose a limit for the high-performance L=0.1μm design point in designing high-speed processors with clock speeds of ~ 500 MHz.

Fig. 4 shows simulated delay versus length for lines scaled down in width W, for differing thicknesses T and insulator heights H between the various wiring layers [as in 5]. The wires studied have neighboring wires separated by a space S, where S=W. Wire tracks above and below are half occupied. All lines have low-impedance drivers, but only one line is driven. The rise time is typical of the fastest possible for this CMOS technology.

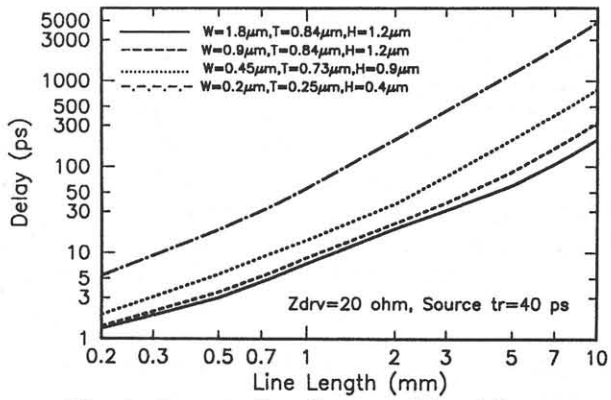

Fig. 4. Impact of scaling on wiring delay.

It is seen that even the smallest line is capable of fast response up to 1 mm in length. The W=0.45μm line is the appropriately scaled middle-level wire to achieve the density shown in Fig. 3 for the L=0.1μm. generation. It has less than 100 ps delay for lines up to 4 mm which is adequate for wiring up control lines in functional blocks.

A hierarchical design and wiring system is needed for high-performance chips where larger dimension wires are used for long data and clock lines. At this stage of scaling, it appears that only the longest lines which already have significant RC delay are inadequate when scaling to smaller dimensions and higher speeds. The remedy is to not scale those lines in width and thickness. This requires either additional levels or some loss of density.

Scaling of the type shown for the W=0.45μm line in Fig. 4, where the space is less than the thickness, can have serious coupling problems. Though it has good RC response, noise of about $V_{DD}/4$ can easily be generated on a long line when its neighbors are driven in the same polarity [5]. Also, it suffers substantial additional delay when

driven in the opposite polarity at the same time as its neighbors. This type of problem exists in present chips, but becomes worse if T and H are not scaled with W.

As systems on a chip become faster and chip sizes stay the same or grow, it is clear that all the problems previously found on multichip modules or boards will then appear on long interconnection lines and power busses of these new chips. Inductance is very important in all such high-speed nets, and careful design of V_{DD} and ground lines is necessary to avoid serious noise coupling problems. In addition, decoupling capacitance on the power busses must be distributed across the chip to supply the peak current demands of the fast-switching circuits.

In terms of the present scaling rules of Table 1, the current density can be seen to scale as $\varepsilon\alpha_w$. Present Al/Ti metallurgy is being used at near maximum current density. Thus higher electric field and smaller wiring dimensions require a better metal such as Cu or Ag, which also can provide lower resistance.

4. SCALING OF FUTURE CMOS DEVICES

Besides the fundamental issue of non-scaling of V_t due to leakage current, several other issues become important in projecting into the sub-0.1μm era. One problem is scaling of the depletion region at the source edge and underneath the gate at threshold. Analytically and empirically this is known to be an important parameter which does not scale because the built-in potential cannot be scaled along with the applied voltage [2]. Changing the doping profile has been useful up to now. Fig. 5 shows the band bending at V_t versus depth under the gate oxide for different design points [6]. The trend from graded to retrograded implant profiles has helped reduce the depletion depth (see dots) almost as much as the device dimensions. A forward bias could reduce the band bending and the depletion depth as shown. This is being used, in a way, in partially depleted SOI devices where the floating body charges to a positive potential due to excess hole generation from various sources.

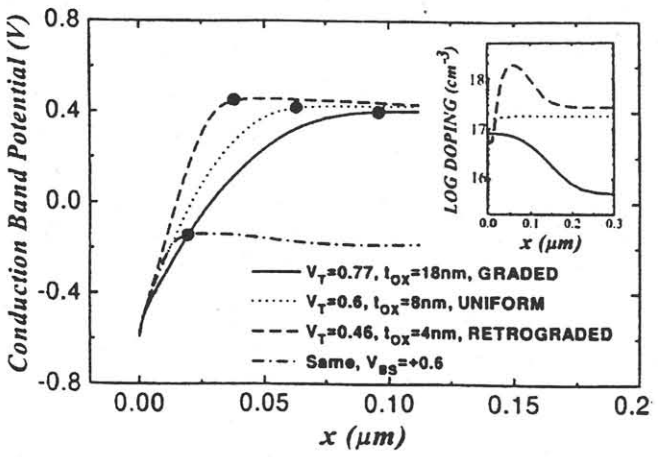

Fig. 5. Scaling of depleted surface at $V_g = V_t$.

Present understanding is that t_{ox} cannot be scaled beyond 2 nm due to tunneling current [7], and no other suitable material exists. Various shorter-channel structures based on thin SOI have been shown to work conceptually, and double-gated devices can make the turn-off sharper so the threshold voltage can be lowered somewhat.

One problem that goes with lower operating voltages is increased sensitivity to soft errors due to ionizing radiation. The amount of charge required to upset a circuit scales down more rapidly than the charge collected from an alpha-particle hit. This is a serious concern for 1.5V and below.

5. DISCUSSION AND CONCLUSIONS

The scaling trends discussed in this paper show that non-scaling of the threshold voltage due to leakage current is already having an impact on future CMOS scaling plans. The increase in power density for high-performance chips will, if chip sizes continue to grow as expected, severely impact the range of applications. However, scaling to lower voltage to maintain better power density will limit the performance benefit. The original constant electric field scaling reduces power-delay product of a given circuit by α^3. If voltage becomes constant, the power-delay product will only improve by α_w with further scaling.

Interconnection problems are becoming significant enough that a new material is needed soon. The possible improvement in resistance (without low-temperature operation) will only solve the problem for 1-2 generations of faster devices. It appears that some sacrifice of the traditional density improvement is necessary to realize the performance benefit of future scaled CMOS.

Thus, there are many challenges to future CMOS scaling and it appears that we are approaching limits to the benefits of further miniaturization. Some areas of change have been identified which can help continue progress.

6. REFERENCES

1. R.H. Dennard, F.H. Gaensslen, H.N. Yu, V.L. Rideout, E. Bassous, and A.R. LeBlanc, IEEE J. Solid-State Circ., **SC-9** (1974) 256.
2. G. Baccarani, M. Wordeman, and R.H. Dennard, IEEE Trans. Electron Devices, **ED-31**, (1984) 452.
3. R.M. Swanson and J.D. Meindl, IEEE J. Solid-State Circuits, **SC-7**, (1972) 146
4. B. Davari and R.H. Dennard, Proc. of IEEE, **83**, (1995) 595.
5. A. Deutsch, et al., IBM J. of Res. and Dev., **39**, (1995) 547.
6. H.J.C. Wann, private communication
7. Y. Taur, et al., IBM J. of Res. and Dev., **39**, (1995) 245.