Circuit-Performance Implications for Double-Gate MOSFET Scaling below 25 nm

Sriram Balasubramanian*, Leland Chang, Borivoje Nikolic and Tsu-Jae King
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA
*Phone: (510) 643-2558, Fax: (510) 643-2636, E-mail: bsriram@eecs.berkeley.edu

ABSTRACT

Circuit-performance implications for double-gate MOSFET scaling in the sub-25 nm gate length regime are investigated. The optimal gate-to-source/drain overlap needed to maximize drive current is found to be different than that needed to minimize FO-4 inverter delay due to parasitic capacitances. It is concluded that the effective channel length must be slightly larger than the physical gate length in order to achieve optimal circuit performance.

INTRODUCTION

The double-gate MOSFET (Fig. 1) is a promising structure for CMOS scaling into the sub-30 nm regime [1]. This structure utilizes a very thin body to eliminate sub-surface leakage paths between the source and drain, and thereby provides excellent control of short-channel effects. The use of a lightly doped or undoped body is desirable for immunity against dopant-fluctuation effects which give rise to threshold-voltage variation, and also for reduced drain-to-body capacitance and higher carrier mobility which provide for improved circuit performance. The threshold voltage of a lightly doped DG-MOSFET is adjusted by tuning the work function of the gate material [2]. In this work, device simulations have been carried out using calibrated energy transport models in MEDICI [3], cross-checked with the quantum device simulator NanoMOS 2.5 [4] for the case of 13 nm gate length. The parameters used in the simulations are summarized in Table 1.

TRANSISTOR DESIGN OPTIMIZATION

The aim of transistor design optimization has typically been to minimize intrinsic gate delay defined as $CV_{dd}/I_{d,sat}$, where C is gate capacitance in inversion, V_{dd} is the supply voltage and $I_{d,sat}$ is the saturation current. At constant V_{dd} and T_{ox}, this translates to maximizing $I_{d,sat}$ (Fig. 4), given a specified maximum leakage current (I_{off}) determined by power dissipation limits [5]. Using this approach, a 25 nm gate length (L_g) DG-MOSFET can be optimized by changing the source-to-drain separation in order to achieve the maximum drive current (Fig. 2). The optimal S/D-separation is determined by the tradeoff between short-channel effects (SCE) and series resistance (R_s). For a fixed gate work function (Φ_M), the leakage current increases as the S/D separation decreases, due to increased SCE (Fig 3). Thus, Φ_M is adjusted in order to meet the I_{off} specification: as the S/D separation decreases, a higher Φ_M is used to compensate for the increased leakage due to increased SCE. When L_g is scaled from 25 nm to 13 nm, the optimal S/D-separation is actually larger than L_g (Fig. 5). This indicates that it will be necessary to employ an effective channel length that is larger than the physical gate length, in the sub-10 nm L_g regime.

EFFECT OF PARASITIC CAPACITANCES

The usual approach for transistor optimization does not take into consideration the role of parasitic capacitances and their impact on circuit performance. The gate capacitance consists of the channel capacitance and parasitic overlap and sidewall capacitances [7]. As the source and drain regions come closer together, the overlap capacitance between the gate and the S/D regions increases (Fig. 6). As the height of the gate electrode is increased, the parasitic sidewall capacitance increases (Fig. 7). Also if raised source/drain (S/D) is employed, the series resistance is reduced, however, there is an additional contribution to the fringing sidewall capacitance (Fig. 7). While a shorter gate height and S/D regions are desirable for achieving lower sidewall capacitance, their heights are usually determined by sheet resistance requirements in order to keep parasitic resistances low. If the gate height and raised S/D regions are not scaled with L_g, parasitic capacitances will cause further relative performance deterioration.

CIRCUIT PERFORMANCE STUDY

To investigate the effect of parasitic capacitance on circuit performance, mixed-mode simulations of fanout-of-4 (FO-4) inverter buffer chains ($W_p/W_n=2$) were carried out using MEDICI. It should be noted that the diffusion capacitance for a thin-body transistor is very small, so that the optimal fan-out can be expected to be close to 3 [7], however, to enable comparison with bulk CMOS reference designs and for lower total power, in these simulations it is increased to 4.

The dependence of FO-4 inverter chain delay on S/D separation was investigated (Fig. 8). As the S/D separation increases, the parasitic overlap capacitances become smaller, and hence the delay decreases. Interestingly, the optimal S/D separation for minimum delay is larger than that determined for maximum transistor drive current. This is because the effect of reducing parasitic capacitance is more significant than that of reducing $I_{d,sat}$. When the S/D separation is increased beyond 1.25×L_g, series resistance limits performance, causing the delay to increase with S/D separation. When a raised S/D structure is introduced, the parasitic series resistance is reduced, resulting in an optimal S/D separation that would be even higher than without raised S/D. The optimal S/D separation corresponding to minimal delay should provide for lower dynamic power consumption, because the parasitic portion of the total switching capacitance is lowered significantly.

CONCLUSION

The optimal gate-to-S/D overlap for maximizing drive current decreases with decreasing gate length, and will be negative (*i.e.* gate-to-S/D underlap will be desirable) for DG-MOSFETs in the sub-10 nm L_g regime. The effect of parasitic capacitance on circuit performance is significant, particularly if the thicknesses of the gate electrode and S/D contact regions are not scaled with the gate length. Therefore, the optimal gate-to-S/D overlap for maximizing circuit performance will be lesser than that needed to maximize drive current.

REFERENCES

[1] D. Frank *et al., IEDM Technical Digest*, p. 553, 1992.
[2] P. Ranade *et al., IEDM Technical Digest*, p. 363, 2002.
[3] MEDICI v2002.4 User's Manual, Avant! Corp., 2000.
[4] NanoMOS 2.5, http://nanohub.purdue.edu
[5] L. Chang *et al., IEDM Technical Digest*, p. 719, 2000.
[6] K. Suzuki *et al., IEEE TED*, Vol. 46, p. 1895, 1993.
[7] J. Rabaey, *Digital Integrated Circuits: A Design Perspective,* 2003
[8] 2001 ITRS, http://public.itrs.net
[9] M.Y.Chang, *J. Appl. Phys.* 82 (6), 1997
[10] Y. Taur, *IEEE EDL*, vol.16, p.136, 1995

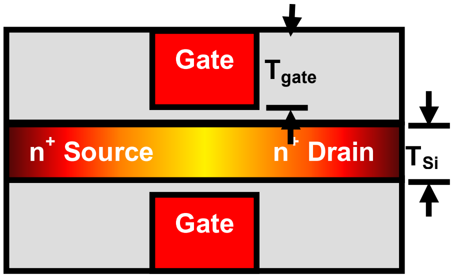

Figure 1: Cross-sectional schematic of the symmetric double-gate transistor structure studied in this work. Graded source-drain doping profiles are used. The structure includes sidewall spacers to capture the effect of parasitic capacitance. The gate work function is adjusted to meet the I_{off} specification. The source and drain contacts are placed on either side.

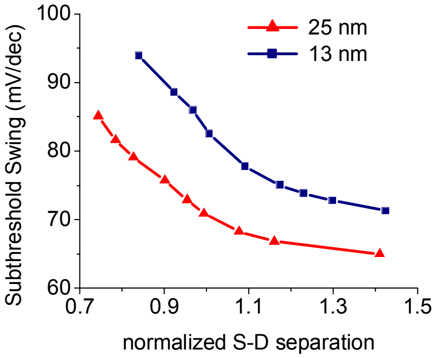

Figure 3: Dependence of sub-threshold swing on S-D separation normalized w.r.t. L_g. As L_g is scaled down, the relative S-D separation required to control SCE will increase.

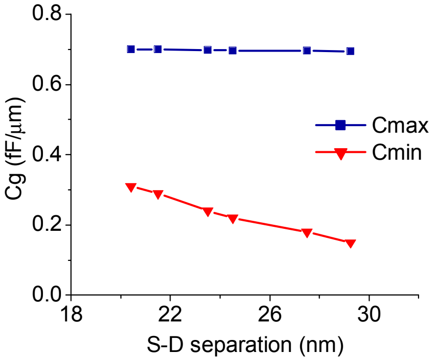

Figure 6: Variation of gate capacitance (C_g) parameters *vs*. S-D separation for L_g=25 nm (assuming a line gate, *i.e.* T_{gate}=0). While C_{max} (@V_{gs}=V_{dd}, V_{ds}=0) remains constant, C_{min} (C_g @V_{gs}=V_{ds}=0) decreases linearly with S-D separation.

Lg (nm)	25	13
T_{ox} (Å)	11	8
T_{Si} (nm)	7	5
V_{dd}(V)	0.7	0.5
Gate height (nm)	37.5	19.5
S-D gradient (nm/dec)	2.8	1.4
S-D doping (cm^{-3})	2x10^{20}	2x10^{20}
$\tau_{relaxation}$ (ps) [9]	1.6	1.3
I_{off} (µA/µm)	0.3	1

Table 1: Parameters used for transistor simulations in this work. These are essentially taken from the ITRS (2001 edition), except that more conservative values of T_{ox} and I_{off} are used.

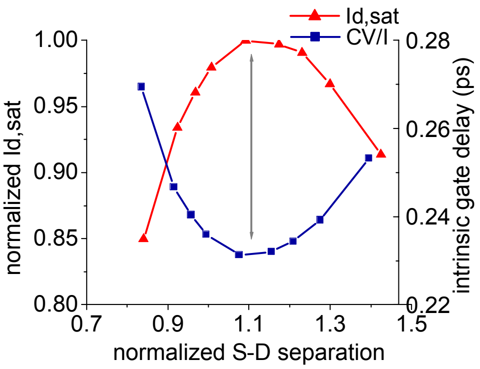

Figure 4: Dependence of $I_{d,sat}$ and intrinsic gate delay (CV/I) on the normalized S-D separation for Lg=25 nm. $I_{d,sat}$ is used henceforth as the metric in place of $CV_{dd}/I_{d,sat}$, since both of them point to the same optimal separation. Since C_{max} remains constant with separation (Fig. 6), minimum CV/I ⇒ maximum $I_{d,sat}$

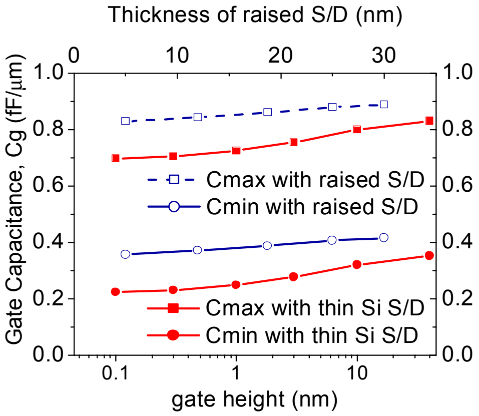

Figure 7: Variation of gate capacitance (C_g) with the gate height (T_{gate}) for L_g=25 nm and 23.5 nm S-D separation. In addition, the effect of fringe capacitance from raised S/D regions is also shown as a function of thickness of raised S/D (shown along the top x-axis) for T_{gate} = 37.5 nm

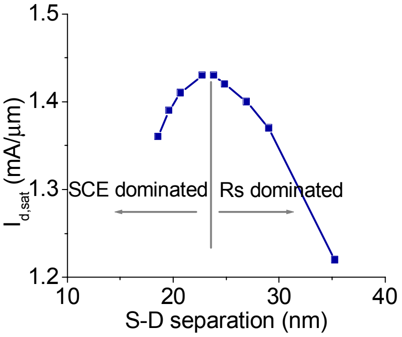

Figure 2: Dependence of $I_{d,sat}$ on source-to-drain separation, for L_g=25 nm. The separation is defined at the positions where the S/D dopant concentration falls to 2x10^{19} cm^{-3} [10]. The gate overlap is symmetric for source and drain.

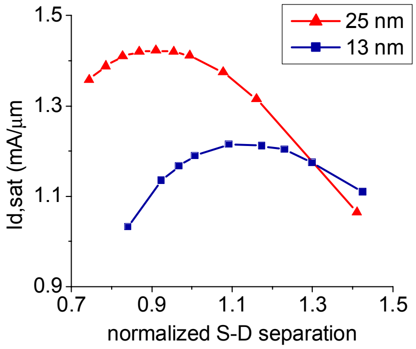

Figure 5: Dependence of $I_{d,sat}$ on the normalized S-D separation. The optimal S-D separation increases as L_g is scaled down, and will be larger than L_g in the sub-10 nm regime.

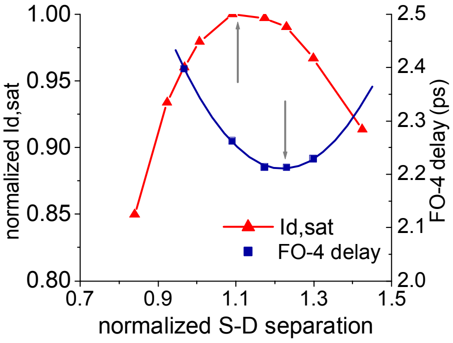

Figure 8: Dependence of $I_{d,sat}$ and FO-4 delay on normalized S-D separation for L_g=13 nm. The optimal separation from the delay perspective is clearly larger than that used to maximize $I_{d,sat}$. The FO-4 simulations were carried out for thin Si S/D.