Trends and Requirements of Future FETs Based on a Simple Physical Device Model

Dimitri A. Antoniadis and Ali Khakifirooz
Microsystems Technology Laboratories, Massachusetts Institute of Technology
60 Vassar St., Cambridge, MA 02139 e-mail: daa@mit.edu

Abstract

A simple analytical model based on seven physical parameters describes the MOSFET in saturation from subthreshold to strong inversion. The model is used to derive a new formulation of the intrinsic switching delay of an FET which is used as a metric of device performance. It is shown that the carrier velocity in the MOSFET channel at the top of the barrier near the source (virtual source) is the main driving force for improved transistor performance with scaling. A historical trend of channel velocity including the most recent results of strain engineering is presented and is used to examine the tradeoffs between key device elements required in order for the performance scaling trend to continue in future "high-performance" CMOS generations.

Performance Metric and Historical Trend

A simple analytical piece-wise model that describes MOSFET $I(V_{GS}, V_{DS})$ in saturation is given below [1]:

$$I_D/W = \min\left[C_{\rm inv}(V_{GS}-V_T)v, V_{DS}/R_T\right] \quad \text{for } I_D/W \geq I_{\rm ref}/L_G \quad (1)$$

$$I_D/W = (I_{\rm ref}/L_G)10^{(V_{GS}-V_T)/S} \quad \text{for } I_D/W < I_{\rm ref}/L_G \quad (2)$$

where $C_{\rm inv}$ is the gate capacitance per unit area at $V_{GS} = V_{DD}$, $V_T = (V_{T0} - \delta V_{DS})$ is the extrapolated saturation threshold voltage [2] which is related to V_{T0} at low V_{DS} and the DIBL parameter $\delta = \partial|V_T|/\partial V_{DS}$, $R_T \approx 2R_S$ is the total S/D and channel resistance, where for simplicity the latter can be assumed negligible, $I_{\rm ref}/L_G$ is the current at $V_{GS} = V_T$ with $I_{\rm ref}$ found empirically to be 3μA and 2 μA for nFETs and pFETs respectively, and S is the subthreshold swing in V/decade. The carrier effective velocity at the virtual source, is given by [1]:

$$v = \frac{v_{x0}}{1 + WR_SC_{\rm inv}(1+2\delta)} \quad (3)$$

where v_{x0} is the actual virtual source velocity, discussed later [1]. All five physics-related parameters: v, V_T, R_S, δ, and S can be extracted from published (measured) data of devices with known (given) $C_{\rm inv}$ and L_G. Based on the model in (1) an analytical expression is derived that relates the intrinsic MOSFET delay to key technology parameters and the effective velocity of carriers in the channel [1], [2]. The intrinsic transistor delay is defined as $\tau = \Delta Q_G/I_{\rm eff}$, where $I_{\rm eff}$ is the effective current [1], [3] and ΔQ_G is the charge difference between the two logic states, including channel and fringing charges. It follows that:

$$\tau = \frac{(1-\delta)V_{DD} - V_T + (C_f^* V_{DD}/C_{\rm inv}L_G)}{(3-\delta)V_{DD}/4 - V_T}\frac{L_G}{v}, \quad (4)$$

where C_f^* is the equivalent fringing capacitance, including Miller effect. The minimum possible value for C_f^* occurs for isoplanar channel and source and drain (S/D), i.e. no raised S/D or contact vias in the vicinity of the gate, and it is roughly 0.5 fF/μm for well-optimized devices and nearly independent of the gate length [4]. Figure 1 shows the historical trend of the intrinsic delay for some benchmark technologies [2].

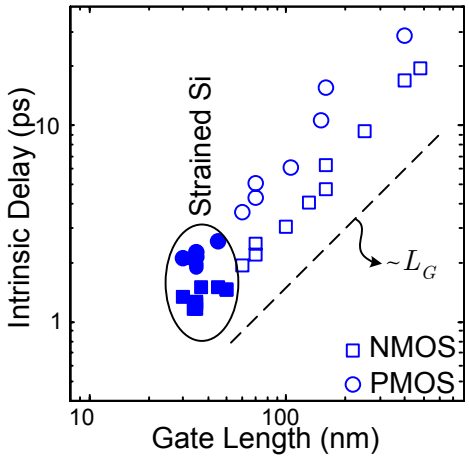

Fig. 1. Historical trend of the intrinsic transistor delay for some benchmark technologies [2]. Filled symbols represent strain-engineered devices. The results signify the fact that strain-engineering is required to continue historical trend of scaling.

Velocity Evolution

Equation (3) shows the relationship between the effective and actual virtual source velocity. This v_{x0} is in-turn related to the ballistic velocity by: $v_{x0} = Bv_\theta$, through the ballistic efficiency, $B = \lambda/(2l+\lambda)$, where λ is the mean free path and l is the critical length for backscattering to the source [5]. Fig. 2 shows the extracted virtual source velocity for the benchmark technologies vs. L_G [2]. As l decreases in proportion to the channel length, the virtual source velocity increases. However, for gate lengths below 130 nm there is saturation in the velocity for relaxed-Si technologies, most likely due to increased Coulombic scattering that results from increased doping necessary to maintain electrostatic integrity. Innovations in strain-engineering have restored the velocity increase by improving mobility and ballistic velocity. For a given technol-

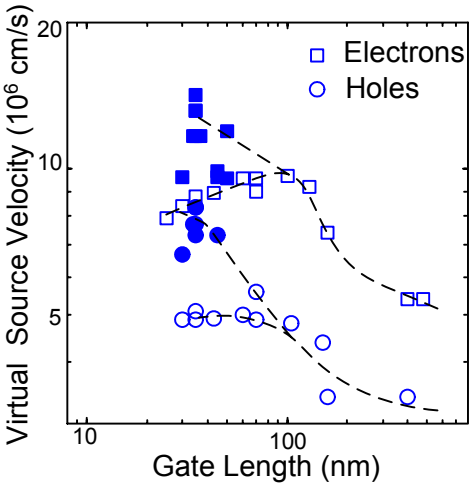

Fig. 2. The extracted virtual source velocity, v_{x0}, vs. gate length for benchmark technologies [2]. Filled symbols represent strain-engineered devices.

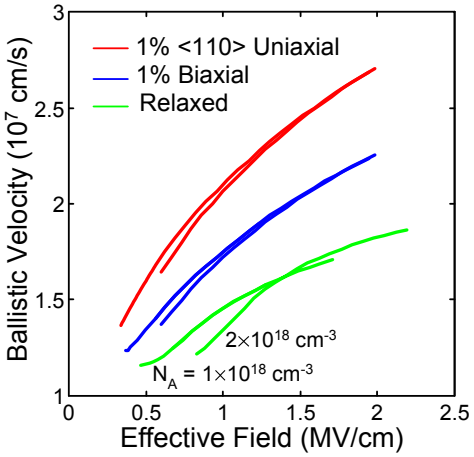

Fig. 3. Calculated electron ballistic velocity vs. effective field for strained and relaxed Si with different doping levels. The change in the effective mass of electrons under uniaxial strain is taken from [8].

ogy carrier velocity increases with decreased electrostatic integrity [6] and therefore, it is important to compare velocities at constant DIBL. As gate length has scaled, the DIBL of the benchmark technologies has increased somewhat, from about 90 mV/V in early technologies, to about 150 mV/V in the most recent technologies. This increase in DIBL has also contributed somewhat in the increase in v_{x0}.

Correlation between Velocity and Mobility

In the degenerate limit, the ballistic velocity is inversely proportional to the square root of both the conduction mass and the density of states mass [7]. Applying mechanical strain increases the velocity by manipulating the effective mass as shown in Fig. 3. Since the mobility is inversely proportional to the effective mass, we can assume that the ballistic velocity is related to mobility via a power law: $v_\theta \propto \mu^\alpha$, where $\alpha \approx 0.5$ when the mobility increase is purely due to the decrease in the effective mass,

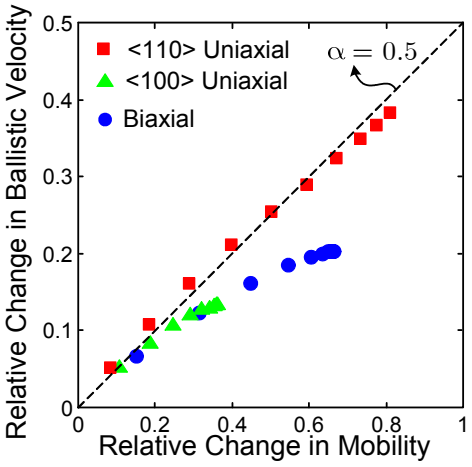

Fig. 4. Relative change in electron ballistic velocity vs. relative change in mobility for different strains calculated based on the data in [8].

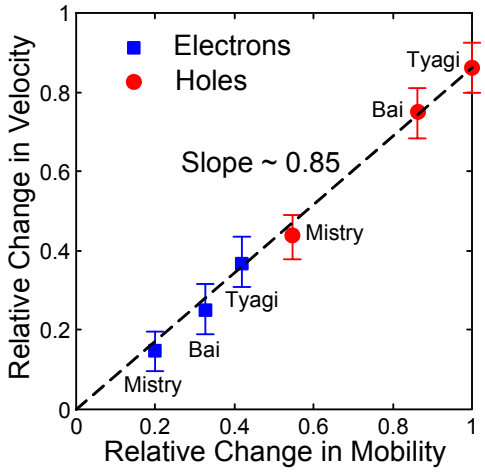

Fig. 5. The relative change in the virtual source velocity vs. the relative change in the mobility based on the data given in [10]. The correlation ratio is much higher than the commonly accepted value of 0.5 [5], [11].

and $\alpha < 0.5$ when reduction in interband scattering is also involved (Fig. 4) [8]. In addition, with more velocity overshoot in the channel, the carrier distribution along the channel drops more abruptly and the electrostatic potential profile is modified to accommodate the change in the carrier distribution [9]. Consequently, l decreases with increasing mobility according to a power law, $l \propto \mu^{-\beta}$, where $\beta \approx 0.45$ [2]. The relative change in virtual source velocity is then related to the change in mobility:

$$\frac{\partial v_{x0}}{v_{x0}} = [\alpha + (1-B)(1-\alpha+\beta)]\frac{\partial \mu}{\mu} \qquad (5)$$

As depicted in Fig. 5 for strain-engineered devices [10], the ratio of the change in velocity to that of mobility is much higher than the commonly accepted value of 0.5 [5], [11].

Eq. (5) provides an alternative method to estimate the ballistic efficiency. Along with the data in Figs. 4-5, it suggests a value of ~0.65 for this parameter. So, these data demonstrate that state-of-the-art MOSFETs operate at ~65% of their ballistic limit. Re-examination of earlier

data in Fig. 2 reveals that unstrained NMOS devices also reached a maximum of about 65% of their ballistic limit before severe Coulomb scattering in heavily doped channels curtailed further increase in B.

Future Directions

With analytical models for device performance, Eq. (4), and off-current, $I_{\rm off} = I_D(V_{GS} = 0, V_{DS} = V_{DD})$ easily obtained from Eq. (2), it is possible to examine the tradeoffs between the seven device parameters of our I-V model for specific goals of performance at a given $I_{\rm off}$. To explain the methodology, the High-Performance (HP) nFET for the 32-nm CMOS generation is examined in some detail. We begin by observing that the key scaling parameter is the so-called contacted gate pitch, $L_{\rm pitch}$, which historically has scaled by 0.7 from generation to generation and is desired to continue to do so.

Table I shows an aggressive scaling scenario for HP CMOS based on this pitch scaling assumption. Numbers in the clear part of the table are taken from the literature along with some personal judgments, while numbers in the shaded part of the table, which refers to future generations are all based on judgments of what is likely technologically feasible. Two additional models are employed: First, the gate capacitance to contact vias is added to the "coplanar" value of C_f^*, with the added assumptions that gate height is equal to $3 \times L_G$, the via half-pitch is equal to the via diameter, and the intervening-medium dielectric constant equal to that of SiN. Second, the "transfer length" formula [12] is used for the calculation of R_S based on some optimistic assumptions about sheet resistance (constant among generations), specific contact resistance (decreasing), and a constant extension resistance $R_{\rm ext} = 45\Omega.\mu$m. In addition, it is assumed that v_{x0} will increase somewhat from the 65 to the 45 nm technology and then stay constant, something that, at least for strain-engineered Si is probably very optimistic, since the decreasing gate pitch will severely limit the space available for stressor materials [13].

Based on the numbers given in Table I, Fig. 7 shows the effect of scaling on gate capacitance and resistance. Note the sharp increase in parasitic capacitance at L_G = 30 nm (45 nm generation) which is due to the decrease in space between gate and contact vias. Also note that the parasitic capacitance has dominated the gate capacitance since approximately L_G =45 nm (90 nm generation). In addition, note that R_S is likely to increase with scaling despite the assumed reduction in specific contact resistance. The result of these effects is that the performance metric, τ, stops scaling at the 65 nm generation and counter-scales thereafter as shown in Fig.8.

From the analytical expression for intrinsic delay, Eq. (4), it is possible to calculate the required parameters in order to continue the downscaling trend of the intrinsic delay with dimensional downscaling at any given generation. A particularly illuminating plot is the required virtual source velocity, v_{x0}, vs. electrostatic integrity, i.e., S and δ, for different source resistance values, R_S. For

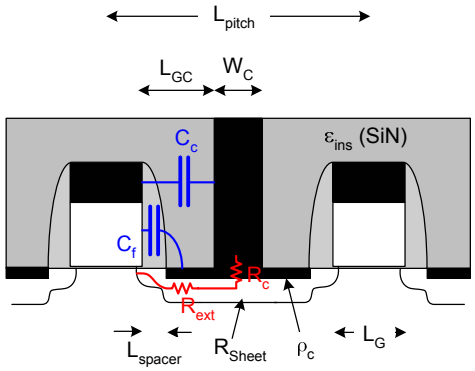

Fig. 6. Illustration of the scaling device features.

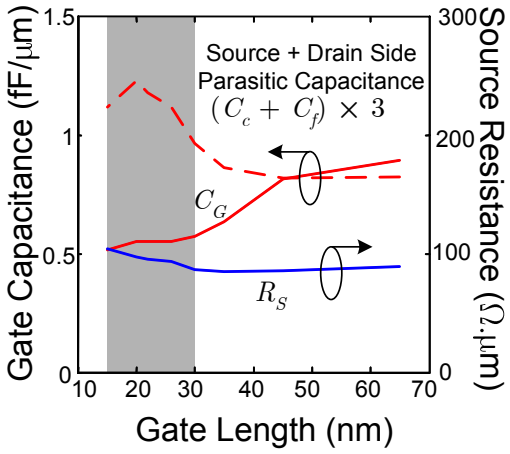

Fig. 7. Gate-channel capacitance, total parasitic fringing capacitance and source resistance vs. gate length for 130 to 15 nm HP CMOS generations. The shaded area highlights projected CMOS generations (45 – 15 nm).

example, referring to Fig. 8, the target delay for the 32 nm generation, obtained by direct extrapolation of intrinsic delay vs. gate length, is 1.1 ps. The results shown in Fig. 9 are quite revealing. As S and δ increase, the required velocity increases drastically because the threshold voltage is forced to ever increasing values in order to maintain the required $I_{\rm off}$. Given that the ballistic efficiency, B, discussed above, is unlikely to exceed 0.7 for well-tempered FETs, the required velocities appear out of reach of strain-engineered Si. This suggests that subject to the device criteria shown in Table I, it will be necessary to seek new channel materials with increased ballistic velocity, v_θ, in order to achieve scalable device performance in the 32 nm generation and beyond. For this reason, it is interesting to evaluate the effect of changing some of the device criteria on the velocity requirements. An example is shown in Fig. 10, where the effect of increasing $I_{\rm off}$ by a factor of 2, or of reducing the gate to contact via capacitance by a factor of about 2, is compared with the default criteria. This also illustrates clearly that the gate stray capacitance is a very important factor in device performance. On the other hand, in the absence of new channel materials, or of reduced gate capacitance solutions, reduction of

TABLE I

HIGH PERFORMANCE (HP) NMOS SCALING SCENARIO

Generation (nm)	130	90	65	45	32	22	15
L_{pitch}(nm)	445	310	220	155	110	84	60
L_G(nm)	65	45	35	30	26	22	15
T_{oxinv} (nm)	2.5	1.9	1.9	1.8	1.63	1.38	1
L_{spacer} (nm)	100	50	30	21	14	10	5
L_{GC} (nm)	125	87	60	40	26	20	15
V_{DD} (V)	1.4	1.2	1.2	1.0	0.9	0.8	0.7
DIBL δ (V/V)	0.10	0.12	0.15	0.15	0.15	0.15	0.15
S (V/dec)	0.12	0.12	0.12	0.12	0.12	0.12	0.12
I_{off}/W (nA/μm)	100	100	100	200	300	300	300
$\rho_c(10^{-8}\Omega.\text{cm}^2)$	6	5	4	3	2.5	2	2
$R_{\text{sheet}}(\Omega/\text{sq})$	250	200	200	200	200	200	200
$v_{x0}(10^7\text{cm/s})$	0.95	1.08	1.38	1.65	1.65	1.65	1.65

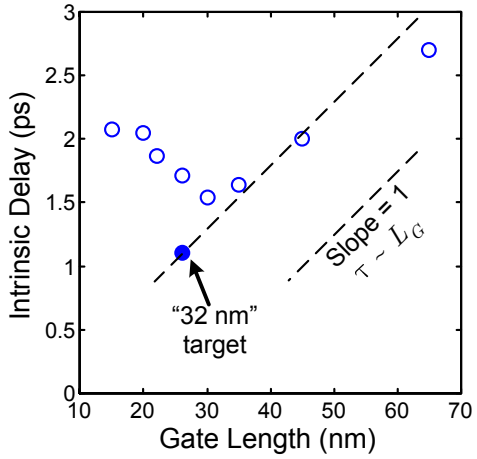

Fig. 8. Intrinsic delay of NFET vs. gate length for 130 to 15 nm HP CMOS generations.

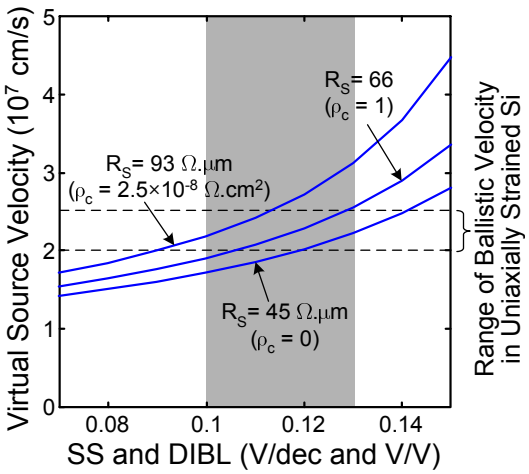

Fig. 9. 32-nm generation nFET virtual-source velocity, v_{x0}, vs. subthreshold swing (S), and DIBL (δ) for different values of silicide-silicon specific contact resistance (ρ_c), required to meet a target intrinsic delay of 1.1 ps. It is assumed that numerically $S = \delta$, which is empirically born out from observations of well scaled modern devices. The shaded region indicates the range of modern device electrostatic integrity. Also shown is the estimated range of maximum achievable ballistic velocity, v_θ, in uniaxially (along [110]) strained silicon channel. Assumed other parameters are: V_{DD} = 0.9 V, I_{off}/W = 300 nA/μm, T_{oxinv} = 1.63 nm, and L_G = 26 nm.

V_T and therefore increase of I_{off} for a limited number of devices in critical paths of the circuit is a partial measure towards increased device performance. However care must be taken in doing this because reducing V_T requires reduction of doping and therefore increase of S and δ. Such tradeoffs can be examined with the introduced analytical models.

Acknowledgments

This work was supported by the MARCO Materials Structures and Devices (MSD) Focus Center.

REFERENCES

[1] D. A. Antoniadis, *et al.*, *IBM J. Res. Dev.*, vol. 50, p. 363, 2006.
[2] A. Khakifirooz and D. A. Antoniadis, *IEDM Tech. Dig.*, p. 667, 2006.
[3] M. H. Na, *et al.*, *IEDM Tech. Dig.*, p. 121, 2002.
[4] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, NY, 1998.
[5] M. Lundstrom, *IEEE EDL*, vol. 22, p. 293, 2001.
[6] H. Hu, *et al.*, *IEEE TED*, vol. 43, p. 669, 1995.
[7] S. Takagi, *Symp. VLSI Tech.*, p. 115, 2003.
[8] K. Uchida, *et al.*, *IEDM Tech. Dig.*, p. 135, 2005.
[9] T. Kobayashi, *et al.*, *IEEE TED*, vol. 32, p. 788, 1985.
[10] S. Tyagi, *et al.*, *IEDM Tech. Dig.*, p. 1070, 2005.
[11] A. Lochtefeld and D. A. Antoniadis, *IEEE EDL*, vol. 22, p. 591, 2001.
[12] S. D. Kim, *et al.*, *IEDM Tech. Dig.*, p. 155, 2005.
[13] J. W. Sleight, *et al.*, *IEDM Tech. Dig.*, p. 697, 2006.

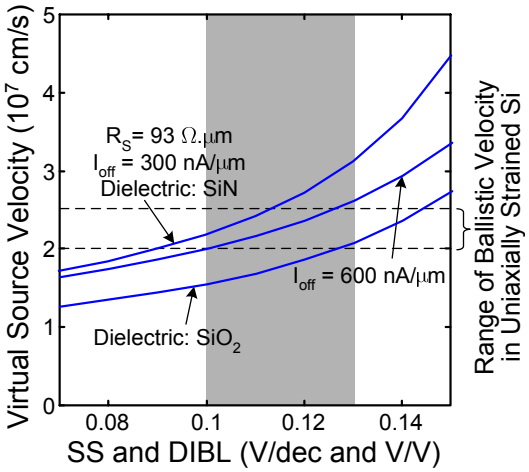

Fig. 10. Same as Fig. 9 except comparing the effect of increasing I_{off} by factor of 2, or reducing the gate stray capacitance to the contact vias by replacing the SiN dielectric by SiO_2 with about 1/2 the dielectric constant.