Monte-Carlo Simulations of Performance Scaling in Strained-Si nMOSFETs

Arvind Kumar[1], Massimo V. Fischetti[2], and Steven E. Laux[1]
[1]IBM Semiconductor Research and Development Center (SRDC), T.J. Watson Research Center,
P.O. Box 218, Yorktown Heights, NY 10598, USA, email: arvkumar@us.ibm.com
[2]Dept. of Electrical and Computer Engineering, Univ. of Massachusetts, Amherst, MA 01003, USA

Abstract — **The performance of strained Si nFETs is studied as a function of channel length and Ge mole fraction using full-band Monte-Carlo simulations. The performance enhancement of strained Si is found to exhibit only modest channel length dependence upon scaling to the 20-nm regime. Although higher Ge mole fraction x leads to an increasing enhancement which is sustained upon channel length scaling, it is argued that x=0.17 represents a good practical design point.**

I. INTRODUCTION

Biaxial strain of Si, realized through epitaxial growth of a pseudomorphic Si layer on relaxed SiGe, is widely considered a promising candidate for achieving performance enhancement. Although enhanced mobility has been clearly observed in long-channel field-effect transistors (FETs) [1,2], with some enhancement reduction measured in 40-nm gate length FETs [3], uncertainty remains about the sustenance of performance enhancements for future technology nodes. Cai et al [4] have experimentally shown only moderate channel length dependence in strained-Si nFETs on SiGe-on-insulator scaled to sub-50-nm channel lengths, provided self-heating is correctly accounted for. The aim of this work is to provide a theoretical basis for understanding the scaling of strained-Si nFETs to the 10-nm regime by studying the dependence of performance on both channel length and Ge mole fraction through full-band Monte-Carlo simulations.

II. DEVICE STRUCTURE

We study the performance enhancement scaling of strained Si by utilizing four bulk n-channel field effect transistors (nFETs) scaled by applying the rules of generalized scaling [5] in which a disproportionately weak decrease of the supply voltage is compensated through additional channel doping. Although in reality the shorter channel length devices would probably be implemented using alternate technologies (*e.g.*, SOI) and high-κ dielectrics, we choose to use simple scaled bulk devices to focus only on the intrinsic performance enhancement scaling of strained Si. As summarized in Table 1, device simulations using the drift-diffusion simulator FIELDAY [6] confirm the proper scaling of device parameters as channel length, junction depth, and gate insulator thickness are decreased; channel doping is increased; and supply voltage is reduced. These scaled devices were then simulated using the full-band Monte-Carlo simulator DAMOCLES [7,8], replacing the unstrained-Si channel with a 13-nm thick strained-Si layer atop a relaxed $Si_{1-x}Ge_x$ buffer, as shown in Fig. 1. Ge mole fractions x of 0 (relaxed), 0.l, 0.17, and 0.25 were chosen to sample the range from weak to strong strain. A (010) surface with current flow along the <100> direction, was assumed.

L (nm)	t_{ox} (nm)	X_j (nm)	N_{ch} (cm^{-3})	V_{dd} (V)	V_{tsat} (V)	SS_{sat} (mV/dec)	DIBL (mV/V)
66	3	5.8	$1.1x10^{18}$	1.5	0.37	88	114
44	2	3.8	$2.2x10^{18}$	1.2	0.32	88	118
22	1	1.9	$6.5x10^{18}$	0.9	0.28	85	112
11	0.5	0.95	$1.6x10^{19}$	0.75	0.12	87	97

Table 1: Basic device input parameters under generalized scaling (gate length L_g, oxide thickness t_{ox}, junction depth X_j, supply voltage V_{dd}) and resulting device characteristics (threshold voltage V_{tsat}, subthreshold swing SS_{sat}, and DIBL).

III. MONTE-CARLO SIMULATIONS

The DAMOCLES simulations utilize full band structure from nonlocal empirical pseudopotentials with spin-orbit coupling. The full band structure of strained Si and $Si_{1-x}Ge_x$ (details given in Ref. [9]) has been used for kinematics and for scattering rates with phonons and ionized impurities. Electron-electron scattering has been turned off to reduce computation time, but a spot check of a few cases found only minor differences when it was included.

In DAMOCLES, interface roughness is treated using a combination of specular and diffuse elastic scattering [8]. Reduced interface roughness scattering is thought to possibly play a role in the enhanced mobility of strained Si

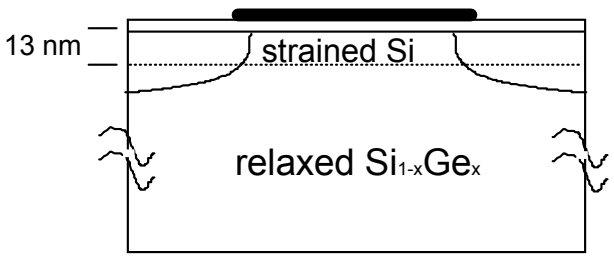


Fig. 1: Bulk nFET device structure used in DAMOCLES simulations consisting of a 13-nm strained Si layer atop a relaxed $Si_{1-x}Ge_x$ buffer. A metal gate with n-type band edge work function was assumed.

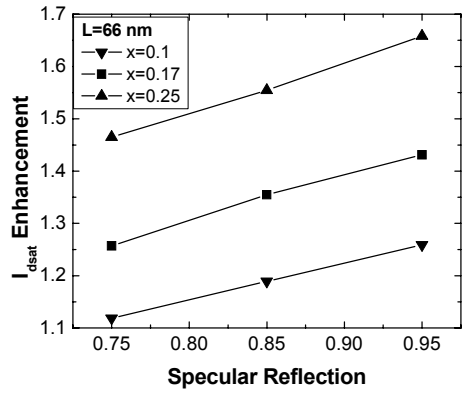


Fig. 2: Strained Si I_{dsat} enhancement as a function of fraction of specular reflection, relative to relaxed Si with 75% specular reflection.

at high transverse fields [10,11]. Figure 2 shows I_{dsat} enhancement as a function of f_{spec}, the fraction of reflection that is specular (i.e., not diffusive), relative to relaxed silicon with f_{spec}=0.75, for L=66 nm. Comparison with experimental data [1] indicates reasonable agreement for drive current enhancement is obtained for f_{spec}~0.75-0.85, and we choose the lower bound value of f_{spec} =0.75 here for simplicity and do not expect the scaling behavior to be strongly affected.

Figure 3 shows the saturation drain current (I_{dsat}), transconductance (G_{msat}), and threshold voltage (V_{tsat}), as a function of gate length L for various x. Because of the practical difficulties of Monte-Carlo simulations with low currents, we compute $V_{tsat}=V_{dd}-I_{dsat}/G_{msat}$ by linear extrapolation in the high gate bias regime at drain voltage V_{dd} where the drain current I_d is linear in V_{gs}. Note that at L=66 nm, V_{tsat} decreases with increasing x, as expected from the bandgap reduction effect, but V_{tsat} is a weaker function of L for increasing x, leading to a different ordering of V_{tsat} in x for small L.

A performance metric that has been found to accurately scale with CMOS inverter delay is the effective drive current $I_{eff}=(I_{low}+I_{hgh})/2$, where I_{low} is I_d for $V_{gs}^{(L)}=V_{dd}/2+\Delta V_t(x,L)$, $V_{ds}^{(L)}=V_{dd}$, and I_{hgh} is I_d for $V_{gs}^{(H)}=V_{dd}+\Delta V_t(x,L)$, $V_{ds}^{(H)}=V_{dd}/2$ [12]. Here $\Delta V_t(x,L)$ refers to the shift in V_{tsat} from the relaxed Si (x=0) case at the same gate length L, allowing us to compare performance at the same gate overdrive. Note that I_{eff} should correlate better with long-channel mobility than an on-current such as I_d for $V_{gs}^{(H)}=V_{dd}+\Delta V_t(x,L)$, $V_{ds}^{(H)}=V_{dd}$, because the I_{low} component is taken at lower drain-source bias.

Figure 4(a) shows I_{eff} as a function of x for different L. Note the acceleration of I_{eff} enhancement for x>0.10, becoming more pronounced at shorter channel lengths. Figure 4(b) shows the I_{eff} enhancement for each x relative to the relaxed (x=0) case. Consistent with experiment [4] and previous Monte-Carlo studies [13], only a weak degradation in performance is observed upon channel length scaling for x=0.17. Below the L=20-nm regime, the x=0.25 case shows

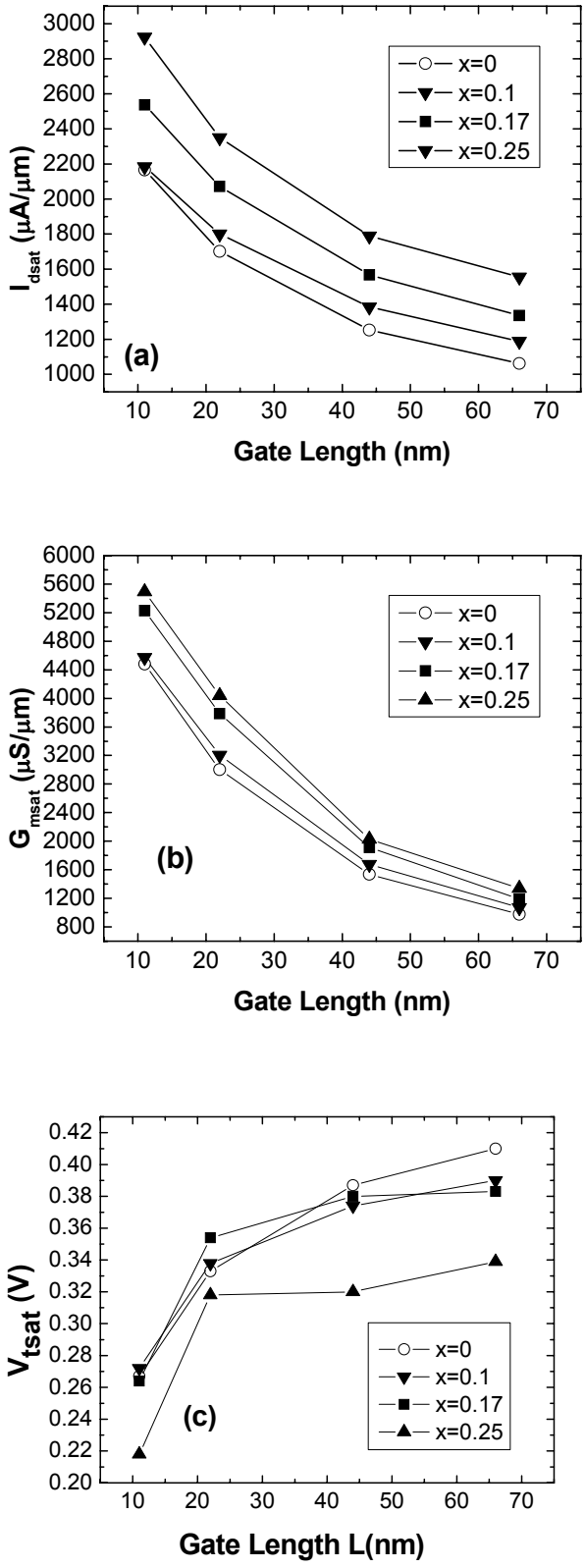
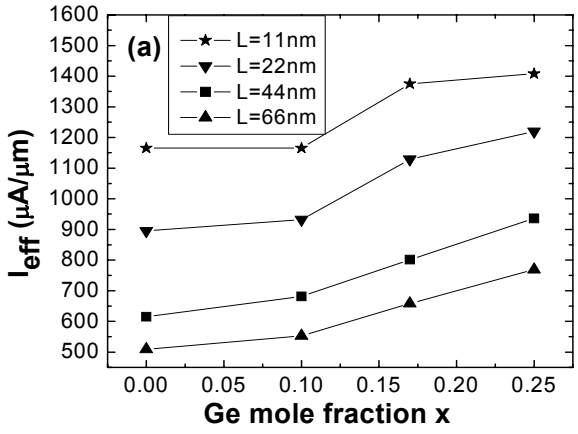
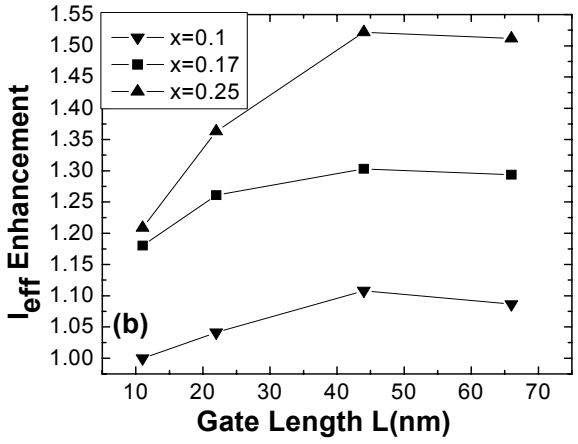


Fig. 3: Saturation (a) drain current I_{dsat} (b) transconductance G_{msat} and (c) threshold voltage V_{tsat} as a function of gate length for various Ge mole fractions.





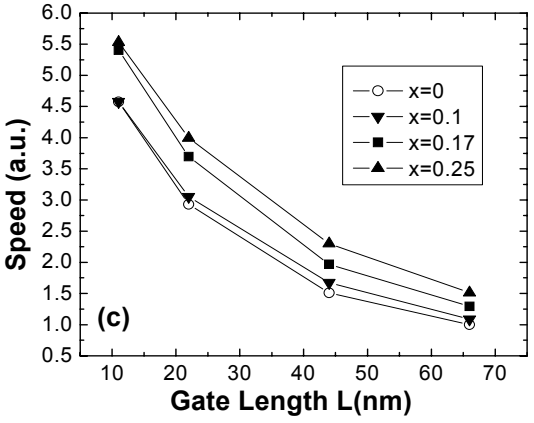


Fig. 4: (a) Effective drive current as a function of Ge mole fraction for different gate lengths. (b) Effective drive current enhancement (normalized to relaxed Si) as a function of gate length for various Ge mole fractions. (c) Scaling of speed with channel length for various Ge mole fractions.

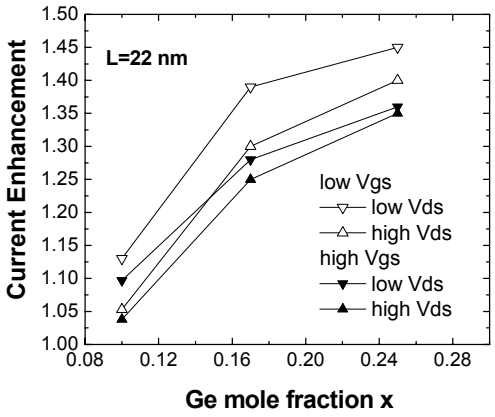


Fig. 5: Current enhancement, relative to relaxed Si, as a function of Ge mole fraction for L=22 nm. For both low and high gate overdrive, a significant fraction of the enhancement seen at low drain bias persists at high drain bias.

substantial degradation, consistent with Ref. [14], and the relatively weak enhancement of the x=0.1 case has disappeared. In Fig. 4(c) we estimate the overall speed scaling based on the dependence of delay as $C_L V_{dd}/I_{eff}$ (load capacitance C_L is assumed to be independent of L). Note the substantial enhancement achieved for x=0.17 as well as its weak channel length dependence. Although intrinsic enhancement continues to increase with x, other considerations -- such as the higher doping that would be required to raise the threshold voltage, higher junction capacitance and leakage, and defect control -- suggest that continuing to increase x is unlikely to be beneficial [15].

An interesting question is the extent to which enhancement at low drain bias correlates to that at high drain bias. At low drain bias, the carriers are in equilibrium and thus the enhancement would be expected to be similar to that seen in the long-channel mobility, whereas performance degradation might be expected to be more pronounced at high drain bias. Figure 5 compares gain at V_{ds} =0.05 V to that at V_{ds} =0.9 V for L=22 nm, for low (V_{gs}= $V_{gs}^{(L)}$) and high (V_{gs}= $V_{gs}^{(H)}$) gate overdrive. Interestingly, even for this short channel length, a significant fraction of the enhancement at low V_{ds} persists at high V_{ds}. This enhancement fraction is seen to increase with increasing x.

Finally Figs. 6(a-b) show the longitudinal velocity along the channel for the L=66 nm and L=22 nm cases, under the I_{high} condition described above. The velocity shown is a weighted average over the electron distribution in the transverse direction. Even for $V_{ds}^{(H)}$=V_{dd}/2, significant velocity overshoot above the saturation velocity is seen. Peak carrier velocity roughly doubles under channel length scaling from L=66 nm to L=22 nm.

IV. CONCLUSIONS

In agreement with recent experimental results, we find using full-band Monte Carlo simulations only modest dependence of strained-Si performance enhancement on channel length down to the 20-nm regime. Sustenance of

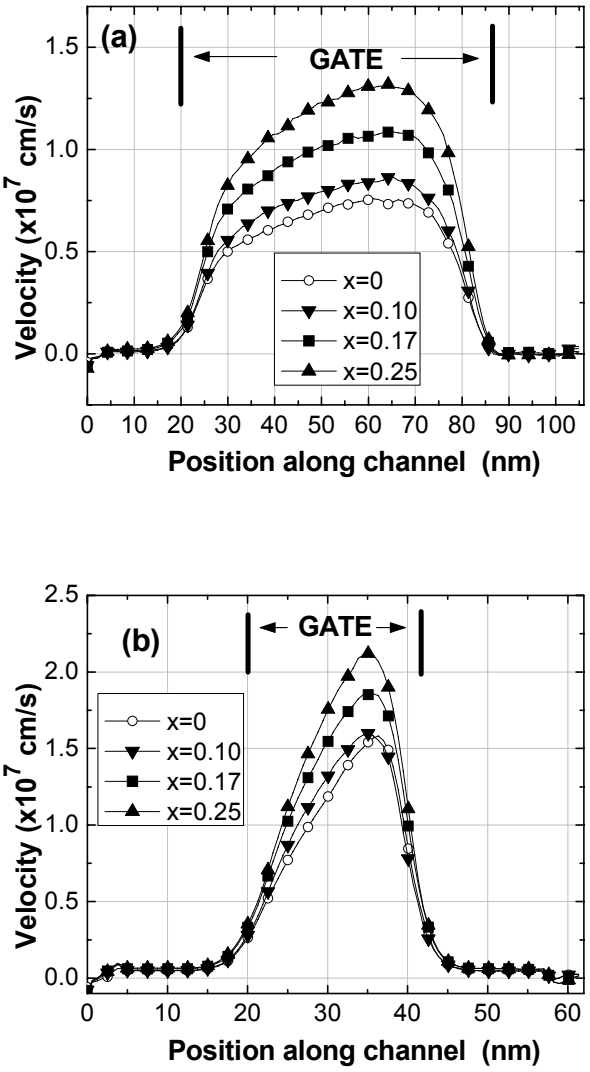


Fig. 6: (a) Density-weighted average of component of electron velocity along channel as a function of position along channel for (a) L=66 nm and (b) L=22 nm, under biasing conditions corresponding to I_{hgh} .

performance enhancement under scaling is confirmed both through persistence of low drain bias gain at high drain bias and through velocity gain arising from overshoot. Intrinsic performance enhancement improves as Ge mole fraction is increased, and these enhancements are robustly sustained under gate length scaling. Ge mole fractions near 0.17 should provide a good design point for two reasons: (1) performance enhancement is well balanced with practical considerations such as junction capacitance and threshold shift, and (2) dependence of performance enhancement on channel length is weaker than for higher strain.

ACKNOWLEDGMENT

The authors wish to thank W. Haensch, P. Oldiges, J. Cai, R. Dennard, and K. Rim for useful discussions.

REFERENCES

[1] K. Rim *et al*., VLSI Tech. Dig., p. 59 (2001).
[2] K. Rim *et al.*, VLSI Tech. Dig., p. 98 (2002).
[3] K. Rim *et al.*, IEDM Tech Dig., p. 43 (2002).
[4] J. Cai *et al.*, IEDM Tech. Dig., p. 165 (2004).
[5] Y. Taur and T. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, New York (1998).
[6] E. Buturla *et al.*, NASECODE VI Proc., p. 291, (1989).
[7] M.V. Fischetti and S.E. Laux, Phys. Rev. B **38**, p. 9721 (1988).
[8] S.E. Laux, M.V. Fischetti, and D.J. Frank, IBM J. Res. Develop. **34**, p. 466 (1990).
[9] M.V. Fischetti and S.E. Laux, J. Appl. Phys. **80**, p. 2234 (1996).
[10] K. Rim *et al*., Solid State Electron. **47**, p. 1133 (2003).
[11] J.R. Watling *et al*., Solid State Electron. **48**, p. 1337 (2004).
[12] M.H. Na *et al.*, IEDM Tech Dig., p. 121 (2002).
[13] F.M. Bufler and W. Fitchner, IEEE Trans. Elec. Dev. **50**, p.278 (2003).
[14] F.M. Bufler, IEDM Tech. Dig., p. 601 (2004).
[15] J.G. Fossum and W. Zhang, IEEE Trans. Elec. Dev. **50**, p. 1042 (2003).