# Impact of Lateral Source/Drain Abruptness on Device Performance

Michael Y. Kwong, *Student Member, IEEE*, Reza Kasnavi, Peter Griffin, James D. Plummer, *Fellow, IEEE*, and Robert W. Dutton, *Fellow, IEEE*

***Abstract*—This paper presents a detailed study of the impact of lateral doping abruptness in the source/drain extension region and the gate-extension overlap length on device performance. Proper choice of the metric used to compare the different device designs is essential. Series resistance and threshold voltage roll-offs are shown to be incomplete measures of device performance that could lead to inconsistent lateral abruptness requirements. While series resistance is seen to improve with increasing junction abruptness, threshold voltage roll-off could be degraded by both lateral junctions that are too gradual and too abrupt—in contrast to the conventional scaling assumptions. The $I_{\rm on}$(supernominal)–$I_{\rm off}$(subnominal) plot, which takes into account statistical variations of gate length, is proposed as a good metric for comparing different device technology designs. Gate-extension overlap length is shown to interact with lateral doping abruptness and to have a significant impact on device performance.**

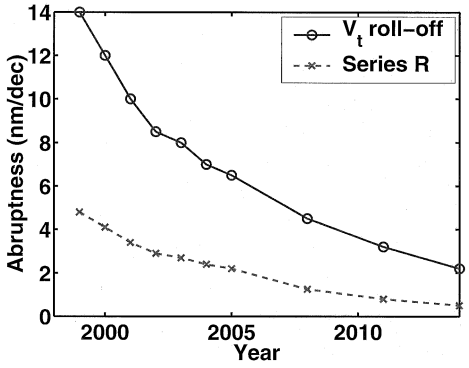



Fig. 1. Lateral source/drain abruptness requirements according to the International Technology Roadmap for Semiconductors based on threshold voltage roll-off and series resistance considerations.

## I. Introduction

DEVICE scaling has resulted in dramatic increases in performance over the past three decades. This is due to the decrease in intrinsic channel resistance as the dimensions of MOS devices become smaller. However, extrinsic series resistance of these devices does not scale well. As a result, it is becoming a significant part of the total device resistance, and its impact on deep submicron device performance can no longer be ignored.

Lateral doping abruptness in the source/drain extension regions is a key parameter that has major impact on source/drain resistance. Ng and Lynch [1] showed the dependence of spreading resistance on lateral abruptness through an analytical calculation. This dependence has been confirmed through a sheet resistance argument [2], examination of the quasi-Fermi level in the extension [3], and rigorous resistance calculations based on two-dimensional (2-D) device simulation [4]. Ghani, *et al.* [5] and Osburn *et al.* [6] showed that the minimal overlap required for proper device operation depends on lateral abruptness, with important implications for device optimization. Furthermore, lateral abruptness has been shown to affect short channel effects [7]. It is obvious that lateral abruptness effects are very important for device design as technology scaling continues.

Fig. 1 shows the requirements for lateral source/drain doping abruptness as predicted by the International Technology Roadmap for Semiconductors (ITRS) [8]. There are two sets of numbers: one based on source/drain resistance [1], and another based on short channel effects [7]. The difference between them is substantial. Before expending resources in trying to achieve the more aggressive targets, it is important to re-examine the basis for these numbers and reconcile the differences.

This paper presents the results of a detailed simulation study of the impact on device performance of lateral doping abruptness and gate-extension overlap length. The methodology used and the important issue of the proper metric for comparing device technologies is discussed in Section II. Then the impact of lateral source/drain abruptness is examined from the viewpoints of series resistance, threshold voltage roll-off and $I_{on}$–$I_{\rm off}$ characteristics (Sections III–IV). Finally, the impact of halo doping, and the assumption that the source/drain extension doping profile is described by a tensor product is examined in Section V.

## II. Methodology

### A. Simulation Details

Device simulation is well suited to the study of lateral abruptness in the source/drain extension. It allows precise control of the doping profiles, which are difficult to ascertain experimentally. The device parameters (Table I) are chosen according to the 2008 technology node on the ITRS roadmap [8]. The 2-D doping profile of the extension is assumed to be a function of the form

$$N(x, y) = N \cdot f_y(y) \cdot f_x(x). \tag{1}$$

To simplify the definition of "lateral abruptness," an exponentially varying doping is assumed at the junctions. Furthermore,

Manuscript received February 25, 2002; revised June 13, 2002. The review of this paper was arranged by Editor J. Vasi.

The authors are with Stanford University, Stanford, CA 94305 USA (e-mail: yiupun@gloworm.stanford.edu).

Digital Object Identifier 10.1109/TED.2002.806790

TABLE I
PARAMETERS FOR SIMULATED DEVICES

| | |
|---|---|
| $V_{dd}$ (V) | 0.9 |
| Lgate (nm) | 50 |
| Poly Doping ($cm^{-3}$) | $5.4 \times 10^{20}$ |
| $t_{ox,eff}$ (nm) | 1.5 |
| Extension $X_j$ (nm) | 20 |
| S/D Extension Peak Doping ($cm^{-3}$) | $2.6 \times 10^{20}$ |
| Substrate Doping ($cm^{-3}$) | $7.5 \times 10^{18}$ |
| Sidewall Spacer (nm) | 40 |
| Contact Size (nm) | 80 |
| Contact Resistivity ($\Omega - cm^2$) | $5 \times 10^{-8}$ |
| Target $V_t$ (V) | 0.35 |

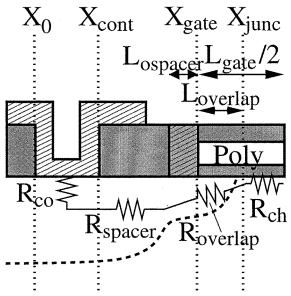



Fig. 2. Schematic of a half MOS device, together with the major resistive components. Also shown is the set of key device parameters that are varied in the simulations.

to help isolate the effects of the lateral abruptness, a uniform channel doping is used. The vertical doping profile of the extension away from the tip is kept constant as the lateral abruptness is varied: an exponentially varying doping of 6.5 nm/dec is used. The doping increases from the junction toward the $Si/SiO_2$ interface until it reaches the "peak value" (analogous to Fig. 5). The lateral abruptness is varied from 1.9 nm/dec to 13.0 nm/dec by changing the lateral to vertical ratio from 0.3 to 2.0. Devices with a range of gate and spacer lengths are simulated with the Lucent mobility model [9]. Note the spacer length determines the amount of gate-extension overlap.

In reality, the source/drain implant would resemble more a Gaussian function, and super-halo profiles [7] will be used, so the channel doping will not be uniform. Nevertheless, the previous simplifications are useful for gaining physical insight. We will revisit these issues in Section VI.

### *B. Metric for Comparing Device Technologies*

Choice of a proper metric for comparing device technologies is critical. The ITRS uses series resistance and threshold voltage roll-off to compare devices with different lateral abruptness.[1] Neither metric by itself is sufficient for comparing device designs. Series resistance does not take $I_{\text{off}}$ into account. Threshold voltage does not fully reflect the impact of series resistance, which limits $I_{on}$. They both provide only a partial picture of the overall device performance. This is the reason for the discrepancy between the two set of ITRS abruptness numbers (Section I).

$I_{on}$–$I_{\text{off}}$ curves provide an alternative, widely used metric. They reflect both power consumption (leakage) and device performance (circuit delay) and thus present a more complete picture of digital device performance. Connelly and Foisy [10] raised two criticisms of the conventional $I_{on}$–$I_{\text{off}}$ curve. First, it assumes gate length is a free parameter that can be used to tune device performance, such as off-current targets. This is not consistent with lithography-limited processes: typically, the target gate length is the starting point; other device parameters[2] are then chosen to maximize the drive current while satisfying constraints such as maximum $I_{\text{off}}$. Therefore, Connelly and Foisy suggest using channel doping instead of channel length as the free parameter to obtain the nominal $I_{on}$–$I_{\text{off}}$ plots.

[1]In the 2001 ITRS, the requirement dictated by series resistance is dropped.

[2]Such as the channel doping.

The second criticism is that the conventional $I_{on}$–$I_{\text{off}}$ curve fails to account for the impact of gate length variations on circuit performance. Connelly and Foisy [10] has shown that the average leakage of a nominal device can be estimated by considering the leakage of a sub-nominal device 0.78 standard deviations shorter than nominal. At the same time, the expected delay of a chain of devices with nominal gate length can be estimated by considering the delay of a super-nominal device 1.6 standard deviations longer than nominal. Plotting the supernominal $I_{on}$ versus the subnominal $I_{\text{off}}$ better predicts circuit performance than either the conventional or the nominal $I_{on}$–$I_{\text{off}}$ curves.

The impact of lateral abruptness will now be examined from these different viewpoints: series resistance, threshold voltage roll-off, and the conventional, nominal and worst case $I_{on}$–$I_{off}$ plots. The first two metrics are incomplete, but provide valuable insights into device operation. The $I_{on}$–$I_{off}$ plots allow us to draw conclusions regarding the effects of lateral abruptness and gate-extension overlap length. This simulation study also provides an excellent example of the application of various $I_{on}$–$I_{off}$ plots.

## III. SERIES RESISTANCE

Fig. 2 shows a schematic diagram of one-half of a typical MOS device and its major resistive components. Note the presence of the "overlap spacer" (of length $L_{ospacer}$), used to tune the gate-extension overlap length ($L_{overlap}$).

Table II shows the resistive components calculated using

$$R_{sh}(x) = \frac{d\phi(x)}{dx} \Big/ I_{ds} \tag{2}$$

$$R_{cont,\,s} = \frac{\phi_{x=x_{cont,\,s}} - V_s}{I_{ds}}$$

$$R_{cont,\,d} = \frac{V_d - \phi_{x=x_{cont,\,d}}}{I_{ds}} \tag{3}$$

where $\phi$ is the quasi-Fermi level obtained from device simulation [4]. Devices with gate lengths of 50 nm, metallurgical overlaps (between the gate and the source/drain extension) of 14 nm, and various lateral abruptness are examined. The overlap resistance improves from 117 $\Omega$ to 64 $\Omega$ as the lateral source/drain abruptness increases from 13.0 nm/dec to 1.9 nm/dec. For devices with very gradual junctions, the lateral slope extends beyond the overlap region, causing an increase

TABLE II
RESISTIVE COMPONENTS FOR DEVICES WITH $L_{gate} = 50$ nm AND VARIOUS LATERAL SOURCE/DRAIN ABRUPTNESS. THE RESISTANCES ARE CALCULATED AS DESCRIBED IN [4] AND HAVE UNITS OF $\Omega/\mu$m

| abruptness $(nm/dec)$ | $R_{co}$ $(\Omega/\mu m)$ | $R_{spacer}$ $(\Omega/\mu m)$ | $R_{ov}$ $(\Omega/\mu m)$ | $R_{chan}$ $(\Omega/\mu m)$ | $I_{on}$ $(\mu A/\mu m)$ | $I_{\rm off}$ $(nA/\mu m$ |
|---|---|---|---|---|---|---|
| 1.9 | 135 | 27 | 64 | 199 | 630 | 51.0 |
| 3.3 | 135 | 27 | 73 | 213 | 572 | 14.5 |
| 4.5 | 135 | 28 | 81 | 220 | 538 | 6.9 |
| 6.5 | 135 | 29 | 92 | 222 | 513 | 4.0 |
| 9.8 | 135 | 32 | 108 | 214 | 505 | 3.5 |
| 13.0 | 135 | 47 | 117 | 201 | 512 | 4.8 |

in $R_{spacer}$ as well. An increase in the on-currents is also expected as the lateral source/drain abruptness increases, and is confirmed by Table II.[3]

The problem with using source/drain resistance for comparing devices with different abruptness is that off-current is ignored. When the lateral abruptness of the junction is increased from 6.5 nm/dec to 1.9 nm/dec, the on-currents improve by 22.7% while the off-currents degrade by over an order of magnitude. Using source/drain resistance as the sole criterion ignores this trade-off, and overestimates the benefits of having an abrupt junction. As a result, the set of abruptness requirement numbers on the ITRS Roadmap [8] based on the series resistance calculations developed by Ng and Lynch [1] are overly stringent.

## IV. THRESHOLD VOLTAGE ROLL-OFF

Taur [7], [11] showed that gradual lateral junctions in the source/drain extension can degrade threshold voltage roll-off. This will now be examined in detail.

Fig. 3 shows the threshold roll-off characteristics of simulated devices with lateral doping abruptness of 1.9, 6.5, and 13 nm/dec. The overlap spacer lengths ($L_{ospacer}$) are adjusted so that devices with the same gate length have the same metallurgical channel length. Other device parameters are chosen as in Table I. The maximum $g_m$ definition of the threshold voltage is used

$$V_{gs,\max gm} = \underset{V_{gs}}{\text{maximize}} \left.\frac{dI_d}{dV_{gs}}\right|_{V_{ds}=V_{d,lin}} \tag{4}$$

$$I_{crit} = I_d|_{\substack{V_{gs}=V_{gs,\max gm} \\ V_{ds}=V_{d,lin}}} \tag{5}$$

$$V_{t,lin} = V_t(V_{ds}=V_{d,lin}) = V_{gs,\max gm} - \frac{I_{crit}}{\left.\frac{dI_d}{dV_{gs}}\right|_{V_{gs}=V_{gs,\max gm}}}. \tag{6}$$

The excessive threshold roll-off observed in the sub-50-nm devices is due to the use of uniform channel doping (no halo doping). For practical devices with these device dimensions, a carefully designed, highly nonuniform doping profile is needed. Nevertheless, the qualitative trends based on uniform doping provide important insights.[4]

[3]The behavior of the devices with the most gradual slopes is affected by threshold voltage roll-off, which will be examined in Section IV.

[4]The issue of halo doping will be revisited in Section VI-A.

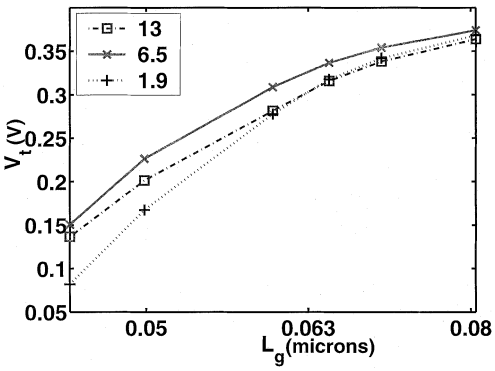



Fig. 3. Plot of threshold voltage ($V_{t,lin}$) versus $L_g$. Note the overlap spacer is adjusted to obtain identical metallurgical channel lengths for devices with the same gate length.

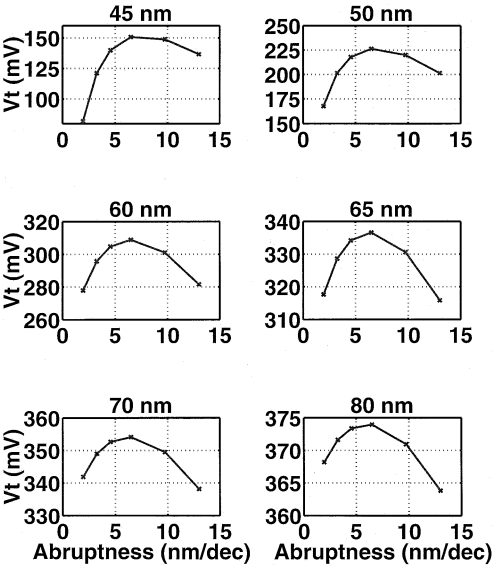
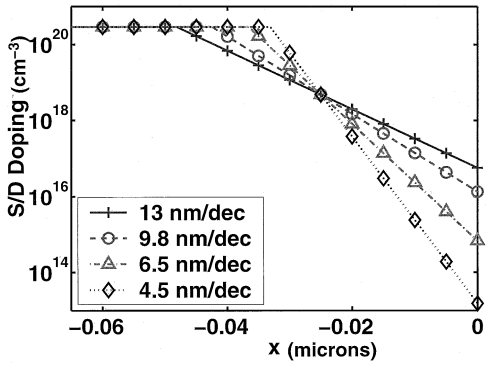



Fig. 4. Plot of $V_{t,lin}$ versus lateral extension abruptness for devices with gate lengths from 45 nm to 80 nm. Note the overlap spacer is adjusted to obtain identical metallurgical channel lengths for devices with the same gate length.

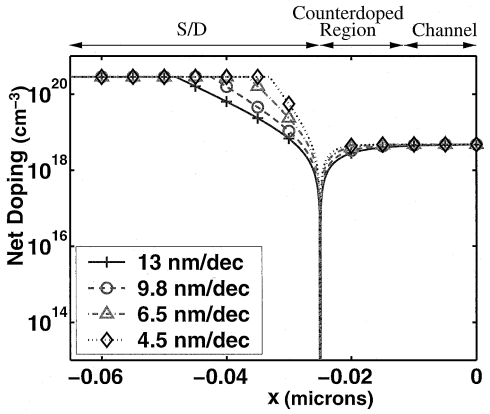
Fig. 4 shows an alternative view of the data. The threshold voltage in the linear region is plotted against lateral source/drain abruptness for devices with the same gate lengths. It is clear that threshold voltage roll-off is degraded by lateral source/drain junction gradients that are too gradual (>6.5 nm/dec), consistent with Crabbé [12] and Taur [7], as well as those that are too abrupt (<3.3 nm/dec), which is contrary to existing literature. This behavior can be explained by considering two competing effects: counterdoping and charge sharing.

### A. Counter-Doping

The degradation of threshold roll-off by very gradual junctions is due to counterdoping of the channel doping [7], [11]. Fig. 5(a) shows four donor doping profiles in devices with lateral source/drain abruptness of 13 nm/dec to 4.5 nm/dec. Fig. 5(b) shows the net doping that results from adding these donor profiles to a uniform channel doping of $4.8 \times 10^{18}$ cm$^{-3}$. The tail of these source/drain donor profiles extends into the channel and counter-dopes the edge of the channel. This in turn decreases the threshold voltage by lowering the net channel doping. The more




(a)




(b)

Fig. 5. (a) Donor doping profile plots along the Si–$SiO_2$ interface, for extensions with lateral abruptness ranging from 13 nm/dec to 4.5 nm/dec. The metallurgical channel length is 44 nm. $x = 0$ is defined to be the center of the channel. (b) Net doping plots along the Si–$SiO_2$ interface, showing counterdoping clearly.

gradual the junction gradient, the more severe counterdoping is, leading to increased degradation in the threshold voltage roll-off. As Taur showed in [11], counterdoping is a direct counterpart of the halo effect [7].

### *B. Charge Sharing*

To understand how junctions with very abrupt lateral gradients can degrade threshold voltage roll-off, we need to consider the origins of short channel effects. Threshold voltage roll-off due to short channel effects can be explained through a charge sharing argument. Unlike the long channel device, a significant portion of the field lines emanating from the bulk charge in a short device terminate in the source and drain regions instead of at the gate. This lowers the threshold voltage. Using the model given by Yau [13], the threshold voltage of an NMOS device is given by

$$V_t = V_{FB} + \frac{\sqrt{-4\epsilon_s q N_a \phi_p}}{C_{ox}} \cdot \left[1 - \frac{r_j}{L}\left(\sqrt{1 + \frac{2x_{d\max}}{r_j}}\right)\right] - 2\phi_p \quad (7)$$

where $L$ is the channel length, $r_j$ is the source/drain junction depth, $x_{d\max}$ is the depletion depth in the substrate, $V_{FB}$ is the flat band voltage of the MOS system and $\phi_p$ is the bulk potential.

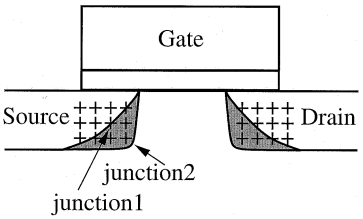



Fig. 6. Device schematic of two source/drain junctions with different lateral abruptness. Junction 2 has the more abrupt doping.

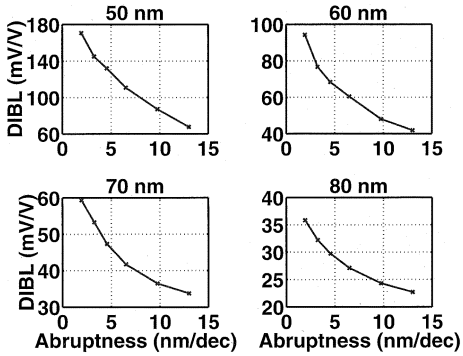



Fig. 7. Plot of DIBL versus lateral source/drain doping abruptness for devices with various gate lengths. DIBL is defined as in (9).

Note that increasing junction depth exacerbates short channel effects.

Fig. 6 shows two source/drain junctions with different lateral abruptness. Assuming the vertical doping abruptness is finite and equation (1) holds, increasing lateral junction abruptness of the source/drain extension will lead to a more box-shaped profile ("junction 2" in Fig. 6). The dopant in the shaded area of the more abrupt device will lead to an increased "effective junction depth" seen from the channel. This will in turn aggravate short channel effects, leading to the degradation in threshold roll-off for the very abrupt junctions shown in Figs. 3 and 4.

Drain induced barrier lowering (DIBL) is another manifestation of the short channel effects. Fig. 7 shows a plot of the amount of drain induced barrier lowering for devices with various lateral abruptness and gate lengths. The threshold voltage at high drain bias is defined using the "critical-current at linear threshold" approach [14]

$$V_{t,\,sat} = V_{gs}|_{I_d = I_{crit},\, V_{ds} = V_{dd}} \quad (8)$$

where $I_{crit}$ is defined in (5). DIBL is then given by

$$DIBL = \frac{V_{t,\,lin} - V_{t,\,sat}}{V_{dd} - V_{dlin}} \quad (9)$$

where $V_{dlin} = 0.05$ V and $V_{t,\,lin}$ is defined in equation (6). It is apparent that the more abrupt junctions exhibit more severe DIBL effects. Note that an increase in (effective) junction depth also degrades DIBL [15].

In summary, there are two problems with ITRS's use of threshold voltage roll-off as a metric [8]. Firstly, threshold voltage roll-off is only a partial measure of device performance and does not take series resistance into account. Secondly, the assumption that increasing abruptness always improves threshold voltage roll-off is not correct.

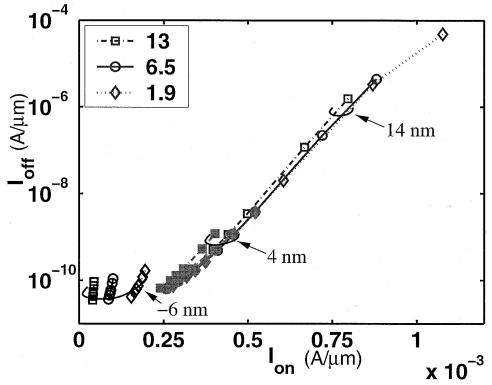



Fig. 8. Conventional $I_{on}$–$I_{\rm off}$ plot for devices with gate lengths from 45 nm to 80 nm, lateral extension abruptness from 13 nm/dec (squares) to 1.9 nm/dec (diamonds), and gate-extension overlap of 4 nm (solid symbols), 14 nm and −6 nm (open symbols).

## V. On-Currents and Off-Currents

We now proceed to examine the use of $I_{on}$–$I_{\rm off}$ plots for examining the impact of lateral abruptness on device performance. These take into account both threshold voltage and series resistance and provide a more complete picture of device performance. The conventional $I_{on}$–$I_{\rm off}$ plots, which are widely used in industry, will be examined first, followed by the refinements discussed in Section II-B.

### A. Conventional $I_{on}$–$I_{\rm off}$ Characteristics

Fig. 8 shows the conventional $I_{on}$–$I_{\rm off}$ plot for several device designs. The free parameter along each curve is the gate length and ranges from 45 to 80 nm. The gate-extension overlap length is varied from 14 to −6 nm, while the lateral gradient of the source/drain extension is varied from 1.9 to 13 nm/dec. For the same target off-current, drive current improves with increasing lateral abruptness. As noted in Section III, the improvement is significantly less than expected from considering source/drain resistance or on-current alone. The improvement in $I_{on}$ by increasing abruptness from 13 nm/dec to 1.9 nm/dec for devices with a constant $I_{\rm off}$ value of 100 nA is less than 5% [Table III(b)], compared with the 22.7% improvement in $I_{on}$ observed in Table II, where $I_{\rm off}$ is not kept constant.

For devices with sufficient gate-extension overlap, as the gate-extension overlap length decreases, the devices stay approximately on the same trend lines. In this regime, decreasing the gate-extension overlap length (while keeping the gate length constant) only serves to increase the metallurgical channel length. Eventually though, the coupling between the channel and the source/drain regions degrades, along with the $I_{on}$–$I_{\rm off}$ performance. This is especially pronounced in devices with the most gradual lateral abruptness, suggesting that devices with abrupt junctions are able to tolerate smaller gate-extension overlap. This is consistent with the results reported by Osburn *et al.* [6].

Note that the conventional $I_{on}$–$I_{\rm off}$ plot assumes the target off-current is achieved by adjusting the channel length. This, as noted in Section II-B, may not be consistent with a typical technology design scenario.

TABLE III
(a) $I_{on}$ (in Units of 0.1 mA/$\mu$m) for Devices With $I_{\rm off} = 0.1$ $\mu$A (or 0.1 nA for $L_g = 70$ nm) and Different Abruptness. (b) Percentage Deviation of $I_{on}$ From the Device With the Best Case Abruptness. Gate-Extension Overlap = 14 nm

| Abruptness | Conv. ($0.1\,A/\mu m$) | Nominal 50 nm ($0.1\,A/\mu m$) | Nominal 70 nm ($0.1\,A/\mu m$) |
|---|---|---|---|
| 13 nm/dec | 6.605 | 6.605 | 3.230 |
| 9.8 nm/dec | 6.746 | 6.743 | 3.298 |
| 6.5 nm/dec | 6.821 | 6.802 | 3.348 |
| 4.5 nm/dec | 6.852 | 6.794 | 3.376 |
| 3.3 nm/dec | 6.871 | 6.755 | 3.394 |
| 1.9 nm/dec | 6.887 | 6.676 | 3.415 |

(a)

| Abruptness | Conv. | Nominal 50 nm | Nominal 70 nm |
|---|---|---|---|
| 13 nm/dec | -4.09% | -2.89% | -5.40% |
| 9.8 nm/dec | -2.05% | -0.87% | -3.40% |
| 6.5 nm/dec | -0.96% | 0.00% | -1.94% |
| 4.5 nm/dec | -0.51% | -0.12% | -1.15% |
| 3.3 nm/dec | -0.24% | -0.69% | -0.61% |
| 1.9 nm/dec | 0.00% | -1.85% | 0.00% |

(b)

### B. Nominal $I_{on}$–Nominal $I_{\rm off}$ Plots

The nominal $I_{on}$–$I_{\rm off}$ plot is obtained with channel doping rather than gate length as the free parameter. As the channel doping is varied, care is taken to maintain the same junction depth and sheet resistance in the source/drain regions by modifying the peak doping. (Note the results do not change substantially if the same source/drain profile is used instead.) Fig. 9 shows the nominal $I_{\rm on}$–$I_{\rm off}$ plots for devices with gate lengths of 70 and 50 nm and various lateral abruptness and gate-extension overlap length. Fig. 9(a) shows that for the 70 nm devices, performance improves with increasing overlap, due to the increased source/drain-to-channel coupling and the shorter channel lengths (for the same gate length).

The behavior of the 50-nm devices is somewhat different [Fig. 9(b)]. While performance still improves with increasing overlap for devices with "underlap," due to the improved source/drain-to-channel coupling, increasing overlap for devices with sufficient overlap does not improve device performance. For ultra short channel devices, short channel effects are severe and dominate. As a result, no further improvements in device performance can be expected from scaling down the channel. Using channel doping as the free parameter in the $I_{on}$–$I_{\rm off}$ plots presents a better picture of the limits of channel scaling for a given technology (given oxide thickness, junction depth and other key design parameters) than the conventional plots.

Table III(a) compares the $I_{on}$ for devices with various lateral source/drain abruptness. In the "conventional case," a channel doping of $4.8 \times 10^{18}$ cm$^{-3}$ is used; the target $I_{\rm off}$ of 100 nA is obtained by tuning the gate length. For the "nominal 50 nm" case, devices with gate lengths of 50 nm are used, and the target $I_{\rm off}$ of 100 nA is obtained by tuning the channel doping. Similarly, in the "nominal 70 nm" case, the target $I_{\rm off}$ of 100 pA is obtained by tuning the channel doping. Note that it is not meaningful to compare the $I_{on}$ obtained for the 50 and the 70 nm de-

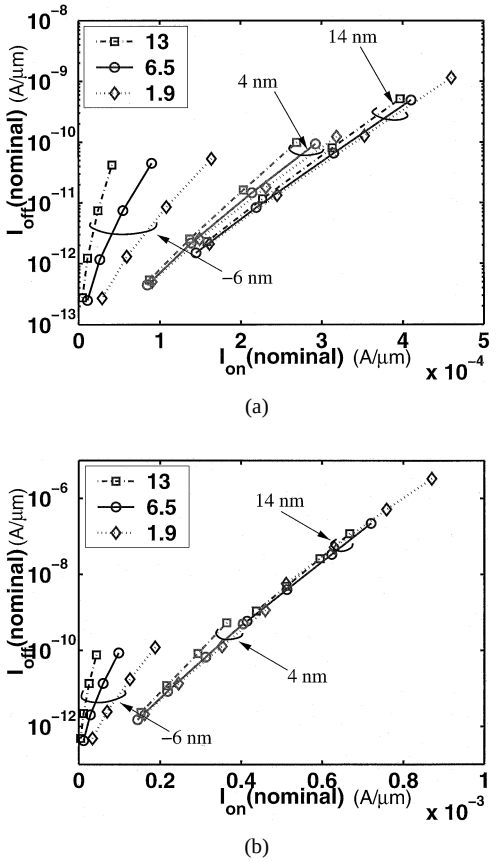



Fig. 9. Nominal $I_{on}$–$I_{\rm off}$ plot for devices with gate lengths of (a) 70 nm (b) 50 nm. Lateral abruptness of the extension varies from 13 to 1.9 nm/dec, gate-extension overlap for the devices from 14 to $-6$ nm, while channel doping varies from 4.8 to 9.0 $\times$ $10^{18}$.

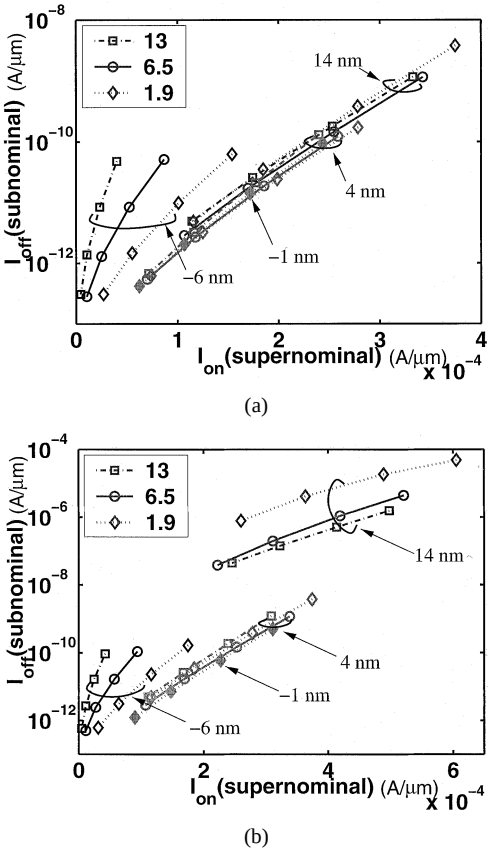



Fig. 10. Plot of supernominal $I_{on}$ versus subnominal $I_{\rm off}$ for devices with nominal gate lengths of (a) 70 nm (b) 50 nm. Gate-extension overlap is annotated on plot and range from 14 to $-6$ nm, lateral extension abruptness from 13 (squares) to 1.9 nm/dec (diamonds), and channel doping from 4.8 to 9.0 $\times$ $10^{18}$ cm$^3$.

vices, since the target $I_{\rm off}$ values are not the same. Instead, the focus is on changes in $I_{on}$ due to varying junction abruptness.

For both the conventional case and the nominal 70 nm devices, increased lateral abruptness improves device performance. The nominal 50 nm devices on the other hand show an optimum abruptness value of 6.5 nm/dec. For these ultra-short devices, short channel effects dominate. The degradation in off-currents resulting from lowered $V_t$ due to the increased lateral abruptness (Section IV) more than compensate for any improvements in $I_{\rm on}$.

### *C. Supernominal $I_{on}$–Subnominal $I_{\rm off}$ Plots*

As discussed in Section II-B, statistical variations of gate length can be accounted for by considering the $I_{on}$ of a supernominal device and the $I_{\rm off}$ of a subnominal one. This turns out to have a major impact on device design.

Fig. 10 shows the supernominal $I_{on}$ versus subnominal $I_{\rm off}$ curves for the same device designs used in earlier discussions. The subnominal device is assumed to have a gate length that is 5 nm shorter than the nominal device, while the supernominal device is assumed to have a gate length that is 10 nm longer than the nominal case.[5]

[5]These values are appropriate for an earlier generation of devices described by Connelly and Foisy [10]. According to the 2001 ITRS, the $L_{gate}$ tolerance for the 50 nm generation is approx 2.5 nm.

Gate-extension overlap length is seen to have a significant impact on device performance. For each gate length and device design, there exists an optimal overlap under this metric. Taking process variations into account highlights the impact of short channel effects, which in turn causes the apparent "optimal" channel length to increase. For short (50-nm) devices, very abrupt junctions with a slight "underlap" have the best performance. The increase in source/drain resistance and the degradation in source/drain-to-channel coupling due to the underlap are more than compensated for by lowered short channel effects due to the increase in metallurgical channel length. The 70-nm devices experience less severe short channel effects. A small overlap of approximately 4 nm is desirable in this case. These results suggest that an "overlap spacer," which allows tuning of the extension overlap, would be useful for device optimization.

It is also shown that the impact of lateral abruptness on device performance depends on the amount of gate-extension overlap. An abrupt junction is desirable for devices with insufficient overlap (the $-6$ nm case), due to the decrease in the source/drain resistance. On other hand, for devices with substantial overlap (the 14 nm case), a more abrupt junction actually hurts device performance, due to increased charge sharing. Note that this is in contrast with Figs. 8 and 9, which does not highlight short channel effects to the same degree.

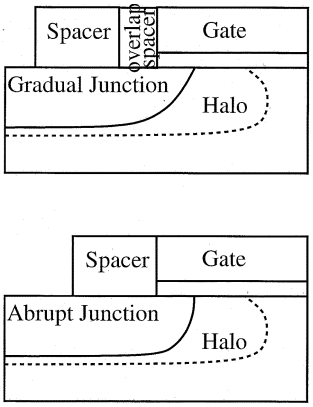



Fig. 11. Overlap spacer is used as a tunable parameter. $L_{met}$ is kept constant for different lateral abruptness by varying overlap spacer length, while $L_{gate}$ remains constant.

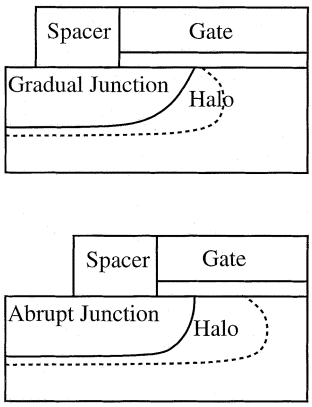



Fig. 12. Overlap spacer is not used as a tunable parameter. $L_{met}$ is kept constant for different lateral abruptness by varying $L_{gate}$.

## VI. Discussion

### A. Halo Doping

Uniform channel doping is assumed for the devices considered thus far. However, for deep submicron devices, halo/pocket implants are required to maintain acceptable short channel behavior [7]. The impact of introducing halo doping on understanding lateral abruptness effects will now be examined. In particular, the focus will be on the threshold roll-off characteristics of devices with different lateral abruptness (analogous to Section IV).

The source/drain doping are described by (1) and the following (similar to [16]):

$$f_y(y) = e^{((y-y_{peak})/\sigma_{y,ext})^2} \tag{10}$$

$$f_x(x) = \begin{cases} e^{((x-x_{peak})/\sigma_{x,ext})^2}: & x > x_{peak} \\ 1: & x \le x_{peak}. \end{cases} \tag{11}$$

The basic source/drain parameters are taken from the well-tempered MOSFET [17]. The metallurgical channel length is 22 nm and $\sigma_{y,ext}$ is 17 nm for all devices. A complete treatment of halo design is outside the scope of this paper, but it is important to note that some halo designs may require tunable spacers to implement. The benefits from the optimal design must be weighed against the complexity in the fabrication process.

Two scenarios are considered. In the first (Fig. 11), the gate length is kept constant at 50 nm. As a result, the location of the halo is fixed. A constant metallurgical channel length for devices with different lateral abruptness is maintained by adjusting the overlap spacer. In the second scenario (Fig. 12), a constant metallurgical channel length is achieved by changing the gate length. As a result, the location of the halo varies for the different devices. Note that the second scenario is the typical scenario considered in the lateral abruptness effects literature.

Fig. 13 shows the threshold voltage and DIBL for devices in the first scenario. The threshold voltage roll-off is degraded by both lateral doping gradients that are too abrupt or too gradual, a result of the interplay between counter doping and charge sharing effects. Also the degradation of DIBL with increased lateral gradient is obvious.

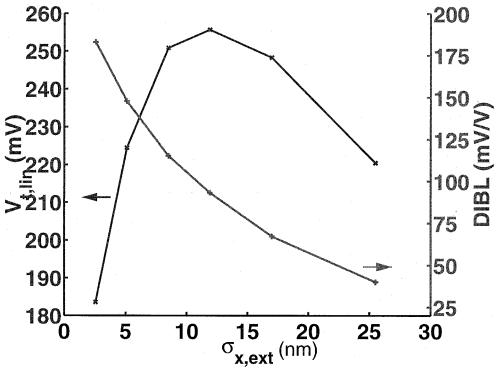



Fig. 13. Threshold voltage and DIBL for devices with fixed halo (Fig. 11). The lateral extension abruptness is controlled by the characteristic length of the Gaussian. Note higher $V_t$ (closer to $V_{t,\,\mathrm{long}} = 0.35$ V) implies decreased threshold roll-off.

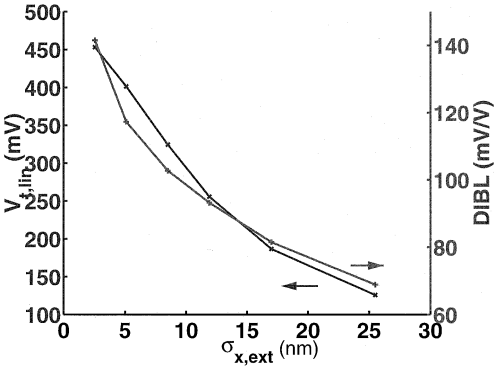



Fig. 14. Threshold voltage and DIBL for devices with varying halo (Fig. 12). The lateral extension abruptness is controlled by the characteristic length of the Gaussian. Note higher $V_t$ (closer to$V_{t,\,\mathrm{long}} = 0.35$ V) implies decreased threshold roll-off/increased threshold roll-up.

Fig. 14 shows the threshold voltage and DIBL for devices in the second scenario. Here the threshold voltage roll-off improves monotonically with increasing junction abruptness (decreasing $\sigma_{x,ext}$). This is consistent with the results in [11], and forms the basis for the claim that short channel effects improve with increasing abruptness. However, DIBL effects are clearly degraded as the lateral gradient increases, suggesting charge sharing effects in fact degrade with increasing lateral junction abruptness.

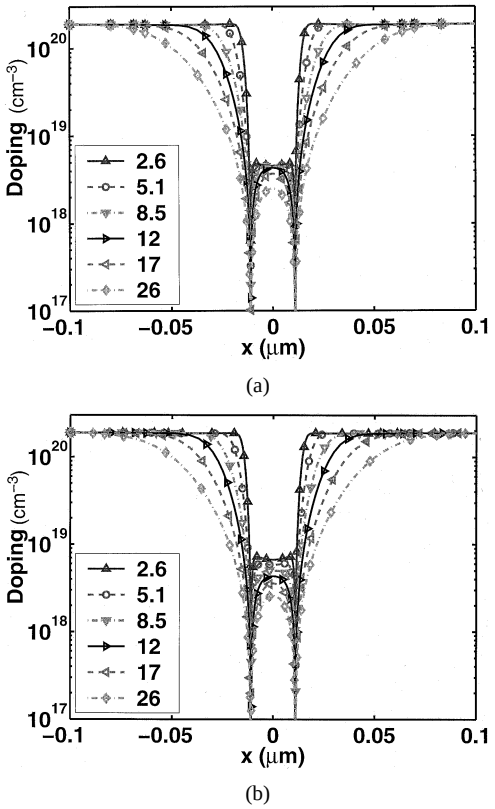



Fig. 15. Doping at the surface of devices in (a) Scenario 1 and (b) Scenario 2, with various lateral abruptness ($\sigma_{x,ext}$ from 2.6 to 26 nm).

The reason for the continual improvements in threshold roll-off as abruptness increases in this scenario is the changing halo location. Fig. 15(a) and (b) show the net doping along the surface. Under scenario two, as the lateral abruptness of the extension increases, the gate length has to be decreased to maintain a constant metallurgical channel length (correlated with the effective channel length), causing the halo to move closer to the center. The doping at the channel center is increased. The increase in the threshold voltage masks the degradation in charge sharing effects, resulting in the threshold voltage plot of Fig. 14. This also explains why the threshold voltage behavior described in Section IV has not been observed by previous publications.

### *B. Source/Drain Doping Description*

Another assumption of this paper is that doping in the source/drain extension region can be described by a tensor product as in (1). Under this assumption, increasing lateral abruptness would lead to a more box-shaped profile. This is a reasonable approximation for doping profiles resulting from ion implantation and diffusion.

However, it is not known if (1) is appropriate for doping profiles resulting from techniques such as laser annealing, a leading candidate for creating ultra abrupt junctions [18], [19]. Privitera *et al.* [20] suggest that laser annealing could result in a box-shaped profile. Hence, using laser annealing as the means to obtain an abrupt lateral doping gradient may also couple an abrupt junction with a box-shaped profile, which would cause charge sharing effects to be degraded. This in turn would lessen the overall benefit of a laterally abrupt junction, as we discussed in Section V.[6]

## VII. Conclusions

Proper choice of performance metrics is important for comparing device design options. Series resistance by itself is not a good criterion for judging device performance, since it considers only on-currents and ignores off-currents. Similarly, threshold roll-off does not take into account series resistance. To reconcile the two sets of ITRS abruptness requirements, both $I_{on}$ and $I_{\rm off}$ must be considered.

For this reason, $I_{on}$–$I_{\rm off}$ curves are useful in comparing different technologies. Two refinements of the conventional $I_{on}$–$I_{\rm off}$ plot are described. Use of the channel doping instead of channel length as the free parameter better matches the typical technology design scenario and quantifies the limits of device scaling. Channel length variations can also be taken into account through consideration of supernominal $I_{on}$ versus subnominal $I_{\rm off}$ device behavior.

A thorough simulation study of the impact of lateral source/drain abruptness and the gate-extension overlap length on device performance is presented in this paper, from the viewpoints of series resistance, threshold voltage roll-off and the three kinds of $I_{on}$–$I_{\rm off}$ plots.

Lateral source/drain abruptness has a dramatic impact on the series resistance of modern MOS devices. This is significant since external resistance makes up a larger fraction of the total resistance of MOS devices as device scaling continues.

At the same time, for source/drain extensions whose doping profile is well described by a tensor product, it is shown that the threshold roll-off characteristics are degraded by a lateral extension junction that is too gradual or too abrupt, and there is an optimal abruptness for a given extension junction depth design, a result of the competition between counterdoping and charge sharing effects. Note that the presence of halo doping could mask this effect.

The gate-extension overlap length is shown to have a significant impact on device performance, suggesting the use of an overlap spacer would be beneficial for device optimization. The impact of lateral abruptness on $I_{on}$–$I_{\rm off}$ depends on the gate-extension overlap length. For devices with substantial underlap, increasing abruptness improves performance. For devices with sufficient overlap, increasing abruptness hurts performance.

In general, increasing lateral source/drain abruptness lowers series resistance, which tends to improve the drive current. However, for very abrupt junctions, this improvement is mitigated by the degradation in leakage currents due to more severe short channel effects. The net improvement in $I_{on}$–$I_{\rm off}$ is less than would be expected if series resistance alone is considered. As a result, the potential improvement in device performance should be carefully weighed against the cost of increasing lateral abruptness.

[6]Increased lateral abruptness is not the only benefit of laser annealing. Laser annealing also increases the vertical abruptness and the activated doping concentration, and allows the construction of ultra shallow junctions [18]. While lateral abruptness by itself may not confer much performance benefits, these factors may combine to offer improved performance for laser annealed devices [16].

## ACKNOWLEDGMENT

The authors would like to thank D. Connelly, Acorn Technologies, Pacific Palisades, CA, M. Duane, Applied Materials, Santa Clara, CA, J. Kluth, C. Riccobene, T. Thurgate, and W. Maszara, Advanced Micro Devices, Sunnyvale, CA, and S. H. Oh, S. Jain, and Y. Takamura, Stanford University, Stanford, CA, for valuable discussions.

## REFERENCES

[1] K. K. Ng and W. T. Lynch, "Analysis of the gate-voltage-dependent series resistance of MOSFET's," *IEEE Trans. Electron Devices*, vol. 33, pp. 965–972, July 1986.

[2] P. Keys, H.-J. Gossmann, K. K. Ng, and C. S. Rafferty, "Series resistance limits for 0.05 $\mu$m mosfets," *Superlatt. Microstruct.*, vol. 27, no. 2–3, pp. 125–136, 2000.

[3] K. Goto, M. Kase, Y. Momiyama, H. Kurata, T. Tanaka, M. Deura, Y. Sanbonsugi, and T. Sugii, "A study of ultra shallow junction and tilted channel implantation for high performance 0.1 $\mu$m pMOSFETs," in *IEDM Tech. Dig.*, Dec. 1998, pp. 631–634.

[4] M. Y. Kwong, C.-H. Choi, R. Kasnavi, and R. W. Dutton, "Series resistance calculation for source/drain extension region with 2-D device simulation," *IEEE Trans. Electron Devices*, vol. 49, pp. 1219–1226, July 2002.

[5] T. Ghani, K. Mistry, P. Packan, S. Thompson, M. Stettler, S. Tyagi, and M. Bohr, "Scaling challenges and device design requirements for high performance sub-50 nm gate length planar CMOS transistors," in *VLSI Tech. Dig.*, 2000, pp. 174–175.

[6] C. M. Osburn, I. De, K. F. Yee, and A. Srivastava, "Design and integration considerations for end-of-the roadmap ultrashallow junctions," *J. Vac. Sci. Technol. B*, vol. 18, pp. 338–345, Jan.–Feb. 2000.

[7] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS design considerations," in *IEDM Tech. Dig.*, Dec. 1998, pp. 789–792.

[8] *International Technology Roadmap for Semiconductors*, Oct. 1999.

[9] M. N. Darwish, J. L. Lentz, M. R. Pinto, P. M. Zeitzoff, T. J. Krutsick, and H. H. Vuong, "An improved electron and hole mobility model for general purpose device simulation," *IEEE Trans. Electron Devices*, vol. 44, pp. 1529–1538, Sept. 1997.

[10] D. Connelly and M. Foisy, "Improved device technology evaluation and optimization," in *Proc. Int. Conf. SISPAD*, Seattle, WA, Sept. 2000, pp. 155–158.

[11] Y. Taur, "MOSFET channel length: Extraction and interpretation," *IEEE Trans. Electron Devices*, vol. 47, pp. 160–170, Jan. 2000.

[12] E. Crabbé, R. Logan, J. Snare, P. Agnello, and J. Sun, "Anomalous short-channel effects in 0.1 $\mu$m MOSFETs," in *IEDM Tech. Dig.*, Dec. 1996, pp. 571–574.

[13] L. D. Yau, "A simple theory to predict the threshold voltage of short-channel IGFET's," *Solid-State Electron.*, vol. 17, pp. 1059–1063, Oct. 1974.

[14] X. Zhou, K. Y. Lim, and D. Lim, "A simple and unambiguous definition of threshold voltage and its implication in deep-submicron MOS device modeling," *IEEE Trans. Electron Devices*, vol. 46, pp. 807–809, Apr. 1999.

[15] R. R. Troutman, "Vlsi limitations from drain-induced barrier lowering," *IEEE Trans. Electron Devices*, vol. ED-26, pp. 461–469, Apr. 1979.

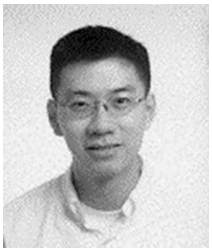
[16] V. Axelrad, A. Al-Bayati, B. Adibi, and P. Carey, "A simulation study of cmos performance improvement by laser annealed source/drain extension profiles," in *Conf. Ion Implantation Technology*, Sept. 2000, pp. 239–242.

[17] D. A. Antoniadis, I. J. Djomehri, K. M. Jackson, and S. Miller. (1999, Aug.) Well-tempered bulk-Si NMOSFET device home page. [Online]. Available: http://www-mtl.mit.edu/Well/.

[18] S. Talwar, G. Verma, and K. H. Weiner, "Ultra-shallow, abrupt, and highly-activated junctions by low-energy ion implantation and laser annealing," in *Proc. Int. Conf. Ion Implantation Technology*, vol. 2, J. Matsuo, G. Takaoka, and I. Yamada, Eds., Kyoto, Japan, Jun. 1998, pp. 1171–1174.

[19] S. B. Felch, D. F. Downey, E. A. Arevalo, S. Talwar, C. Gelatos, and Y. Wang, "Sub-melt laser annealing followed by low-temperature RTP for minimized diffusion," in *Conf. Ion Implantation Technology*, Sept. 2000, pp. 167–170.

[20] V. Privitera, C. Spinella, G. Fortunato, and L. Mariucci, "Two-dimensional delineation of ultrashallow junctions obtained by ion implantation and excimer laser annealing," *Appl. Phys. Lett.*, vol. 77, pp. 552–554, July 2000.



**Michael Y. Kwong** (S'93) received the B.S. and M.S. degrees in electrical engineering from Stanford University, Stanford, CA, in 1995 and 1997, respectively. He is currently pursuing the Ph.D. degree at the Center for Integrated Systems, Stanford University, Stanford, CA.

His research interests include device scaling, doping characterization optimization, and solution of inverse problems.

Mr. Kwong is a member of the IEEE Computer Society.

**Reza Kasnavi** received the B.S. degree from Sharif University of Technology, Tehran, Iran, in 1995, and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA in 1997 and 2001, respectively.

His research interests include process integration and scaling of MOS transistors, compound semiconductors, and bio-mems.

Dr. Kasnavi is a member of the IEEE Electron Devices and Communications Societies, the Materials Research Society, and the Optical Society of America.

**Peter Griffin** received the B.E. and M.E. degrees from University College, Cork, Ireland, in 1981 and 1983, respectively, and the Ph.D. degree from Stanford University, Stanford, CA, in 1989.

His research interests include process integration and scaling of MOS transistors, compound semiconductors, and bio-mems.

**James D. Plummer** (M'71–SM'82–F'85) was born in Toronto, ON, Canada. He received the B.S. degree from the University of California, Los Angeles, and the M.S. and Ph.D degrees in electrical engineering from Stanford University, Stanford, CA.

He is currently the John Fluke Professor of Electrical Engineering, the Frederick E. Terman Professor of Engineering, and Dean of the School of Engineering at Stanford University. He has authored or coauthored over 300 technical papers. His current research interests focus on silicon devices and technology. He is particularly interested in the limits of silicon devices and technology, new application areas for chips, and in exploring possible replacement technologies for silicon chips. He consults for and serves on the boards of a number of semiconductor companies.

Dr. Plummer has received three Best Paper Awards at the International Solid State Circuits Conference. In 1991, he received the Solid State Science and Technology Award from the Electrochemical Society. He has also received several teaching awards at Stanford University. He was elected to the National Academy of Engineering in 1996, and recently received the Semiconductor Industry Association's 2001 University Research Award.

**Robert W. Dutton** (F'84) received the B.S., M.S., and Ph.D. degrees from the University of California, Berkeley, in 1966, 1967, and 1970, respectively.

He is Professor of electrical engineering, Stanford University, Stanford, CA, and Director of Research in the Center for Integrated Systems. He has held summer staff positions at Fairchild, Bell Telephone Laboratories, Hewlett-Packard, IBM Research, and Matsushita during 1967, 1973, 1975, 1977, and 1988, respectively. His research interests focus on integrated circuit process, device, and circuit technologies, especially the use of computer-aided design (CAD) in device scaling and for RF applications. He has published more than 200 journal articles and graduated more than four dozen doctorate students.

Dr. Dutton was Editor of the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN (1984-1986), winner of the 1987 IEEE J. J. Ebers and 1996 Jack Morton Awards, winner of the 1988 Guggenheim Fellowship to study in Japan, and was elected to the National Academy of Engineering in 1991. He was also honored with the C&C Prize (Japan) in 2000.