Analysis of Nonuniform ESD Current Distribution in Deep Submicron NMOS Transistors

Kwang-Hoon Oh, *Student Member, IEEE*, Charvaka Duvvury, *Senior Member, IEEE*, Kaustav Banerjee, *Member, IEEE*, and Robert W. Dutton, *Fellow, IEEE*

***Abstract*—This paper presents a detailed study of the nonuniform bipolar conduction phenomenon under electrostatic discharge (ESD) events in single-finger NMOS transistors and analyzes its implications for the design of ESD protection for deep-submicron CMOS technologies. It is shown that the uniformity of the bipolar current distribution under ESD conditions is severely degraded depending on device finger width (W) and significantly influenced by the substrate and gate-bias conditions as well. This nonuniform current distribution is identified as a root cause of the severe reduction in ESD failure threshold current for the devices with advanced silicided processes. Additionally, the concept of an intrinsic second breakdown triggering current (I_{t2i}) is introduced, which is substrate-bias independent and represents the maximum achievable ESD failure strength for a given technology. With this improved understanding of ESD behavior involved in advanced devices, an efficient design window can be constructed for robust deep submicron ESD protection.**

***Index Terms*—Deep-submicron CMOS, electrostatic discharges, ESD protection, nonuniform electrostatic discharge current distribution, single-finger NMOS.**

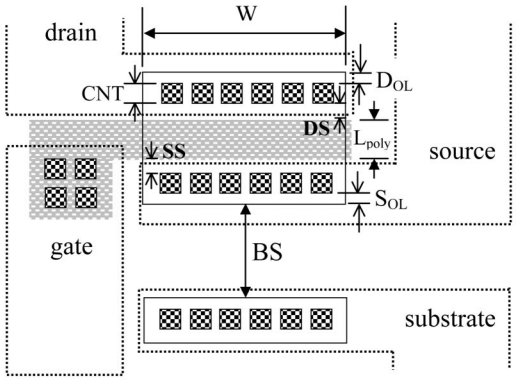


Fig. 1. Layout of the single-finger NMOS transistor. The contact opening (CNT) is 0.15 μm for both 1.5-V and 3.3-V devices. The gate to source/drain contact spacings (SS/DS) are 0.1 μm and 0.225 μm and n^+ overlaps of source/drain contact (S_{OL}/D_{OL}) are 0.4 μm and 0.125 μm for the 1.5-V and 3.3-V transistors, respectively. Also, the body space (BS) from the source diffusion to the substrate diffusion is the same as the finger width (W).

NOMENCLATURE

β	Current gain of the parasitic n-p-n transistor.
f	Frequency of applied pulse in EMMI tests.
I_D	Injected drain current.
$I_{\rm gen}$	Avalanche-generation current.
I_{sp}	Snapback holding current.
I_{t1}	Triggering current.
I_{t2}	Second breakdown triggering current.
I_{t2i}	Intrinsic second breakdown triggering current.
L_{poly}	Drawn gate poly length.
R_s	Series resistance in high current region.
T_p	Pulse width in EMMI tests.
$T_{\rm exp}$	Exposure time in EMMI tests.
V_{t1}	Triggering voltage.
V_{t2}	Second breakdown triggering voltage.
V_{sp}	Snapback holding voltage.
V_{sub}	External substrate-bias.
W	drawn gate poly finger width.
W_{eff}	effective finger width.
$W_{\rm max}$	maximum turned-on finger width.

Manuscript received February 20, 2002; revised July 10, 2002. This work was supported by Texas Instruments Inc., Dallas, TX. The review of this paper was arranged by Editor C.-Y. Lu.

K.-H. Oh and R. W. Dutton are with the Center for Integrated Systems, Stanford University, Stanford, CA 94305 USA (e-mail: okhoon@gloworm.stanford.edu; dutton@ee.stanford.edu).

C. Duvvury is with Silicon Technology Development, Texas Instruments, Dallas, TX 75243 USA (e-mail: c-duvvury@ti.com).

K. Banerjee is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: kaustav@ece.ucsb.edu).

Digital Object Identifier 10.1109/TED.2002.805049

I. INTRODUCTION

FOR ON-CHIP ESD protection circuits, the first requirement for achieving good protection structures is to provide a low-impedance discharging current path to shunt ESD currents and clamp the I/O pad voltage to a safe level without causing damage to internal circuits. The size of protection devices is mainly determined based on the current handling capability required in the ESD protection circuits. The multifinger structures are the most common way of designing the protection devices in various sizes. In multifinger structures, uniform triggering across all the fingers is important to achieve maximum current handling capabilities. This requires that the single fingers in the multifinger structures are fully turned on under ESD conditions.

Typically, multifinger gate-grounded NMOS devices are widely used as protection structures owing to the effectiveness of parasitic lateral n-p-n bipolar transistors in handling high ESD currents. Nonuniform triggering behavior of lateral n-p-n transistors was first reported by Scott *et al.* for silicided single-finger NMOS devices under ESD stress [1] and subsequently the phenomenon of nonuniform triggering in multifinger NMOS transistors and the implications for the design of ESD protection was discussed by Polgreen *et al.*

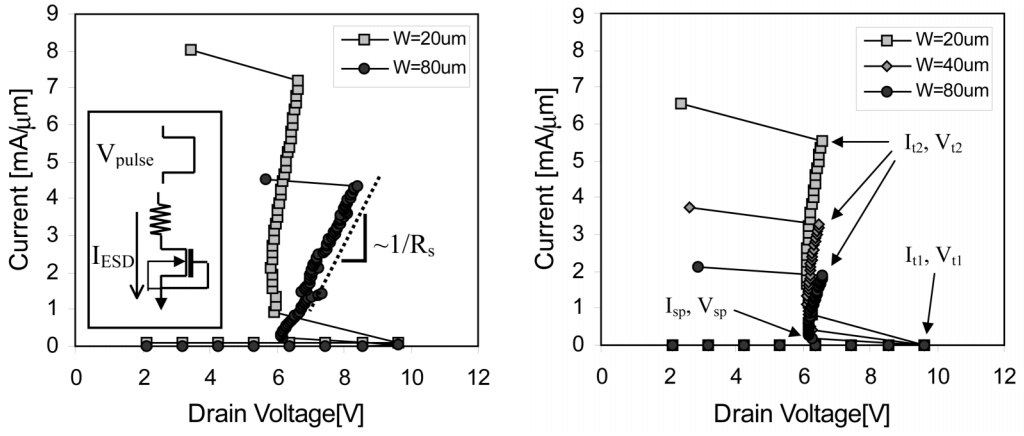


Fig. 2. Sample high current I–V curves for 3.3-V NMOS transistors with different finger widths for (a) nonsilicided and (b) silicided processes where $L_{poly} = 0.5\ \mu$m. The current per unit finger width at second breakdown (I_{t2}) strongly depends on the finger width (W), which illustrates nonuniform bipolar currents flow under ESD conditions.

[2]. The simultaneous triggering of the multifinger NMOS protection structures has been considered as a critical aspect for the effectiveness of ESD protection designs [3]–[5]. However, in recent years the dependence of the second breakdown triggering current (I_{t2}), which is widely used for monitoring ESD robustness of a protection structure, on the single-finger device width (W) has become quite an important criterion for optimum design of ESD protection circuits, since under ESD events current localization occurs in advanced silicided single-finger transistors [6], [7]. We have recently addressed this issue for the designs of protection devices involving advanced submicron technologies [8]. In this study, detailed experimental investigations into the nonuniform current conduction phenomenon have been performed including transmission line pulsing (TLP) measurements and emission microscopy (EMMI) analysis for various silicided and nonsilicided test structures. Moreover, the impact of both substrate and gate-bias conditions on this nonuniform current distribution and their implications for the design of ESD protection have been discussed in detail, which provides new physical insight into the ESD behavior of advanced protection devices and useful basis for constructing efficient design windows for robust ESD protection design to overcome ESD failures in advanced deep submicron technologies.

II. Experimental Evidence of Nonuniform Bipolar Conduction

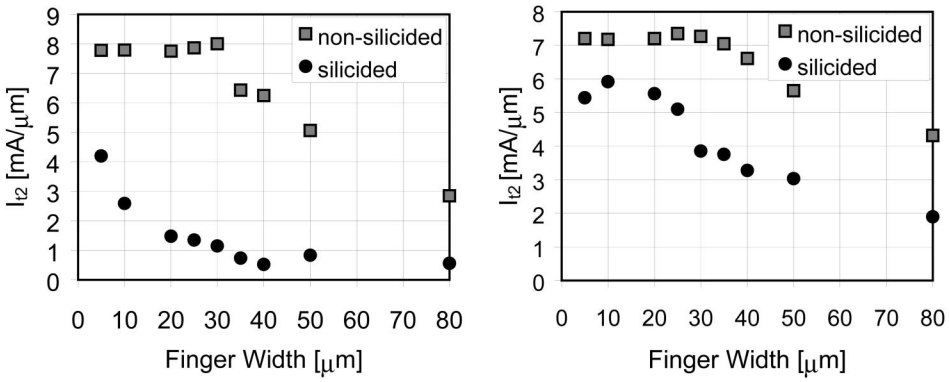
Various test structures, including the low and high-voltage NMOS transistors (1.5-V NMOS with 27-Å-thick gate oxide and 3.3-V NMOS with 70-Å-thick gate oxide) with shallow trench isolation (STI) in a 0.13-μm CMOS technology were investigated in this work. Both silicided ($CoSi_2$) and nonsilicided devices, which were formed on a 3.5-μm-thick epilayer, were explored for comparison. The drawn polygate length (L_{poly}) was 0.175 μm and 0.5 μm for the low-voltage and high-voltage transistor, respectively, and the substrate contact was located parallel to the source contact to keep the substrate resistance constant along the finger width direction as shown in Fig. 1.

A. Transmission Line Pulsing (TLP) Tests

The high current behavior of protection devices can be analyzed by applying a short time-scale constant-current pulse, generated using a transmission line, to protection structures with increased pulse magnitude at each step [9]. Using a TLP system, I_{t2} measurements were performed with a 200 ns long voltage pulse for various test structures and sampled high current-voltage (I–V) snapback curves for the 3.3-V silicided and nonsilicided devices are shown in Fig. 2. As shown in Fig. 2, the single-finger device shows a significant width dependence of I_{t2} as measured in mA/μm and the effect becomes more apparent with silicide process. According to the I–V curves, the application of silicided technology shows no significant differences in the n-p-n transistor triggering voltage (V_{t1}), since V_{t1} is dominantly determined by the avalanche multiplication process across the drain-substrate junction and both devices were drain engineered in the same manner. In addition, the snapback holding voltage V_{sp} of both the silicided and nonsilicided transistors is nearly the same, which implies that the presence of silicide diffusion over the source/drain has no observable impact until the parasitic lateral n-p-n transistor snaps back. However, the series resistance $R_s[\Omega\text{-}\mu\text{m}]$ in the high current regime for silicided and nonsilicided devices shows considerable differences as expected. Those high current I–V curves indicate that the current flows nonuniformly along the finger width (W) after the lateral n-p-n transistor snaps back. For both the low and high-voltage devices, only a limited portion of the finger is effective for ESD current conduction beyond I_{t2} roll-off points as is apparent from the channel width dependence shown in Fig. 3. The width dependence of I_{t2} is a clear evidence of the strong nonuniformity of bipolar conduction, even in single-finger NMOS transistors. Moreover, for advanced silicided technologies, the degree of its severity is even more significant in the effective design of ESD protection circuits.

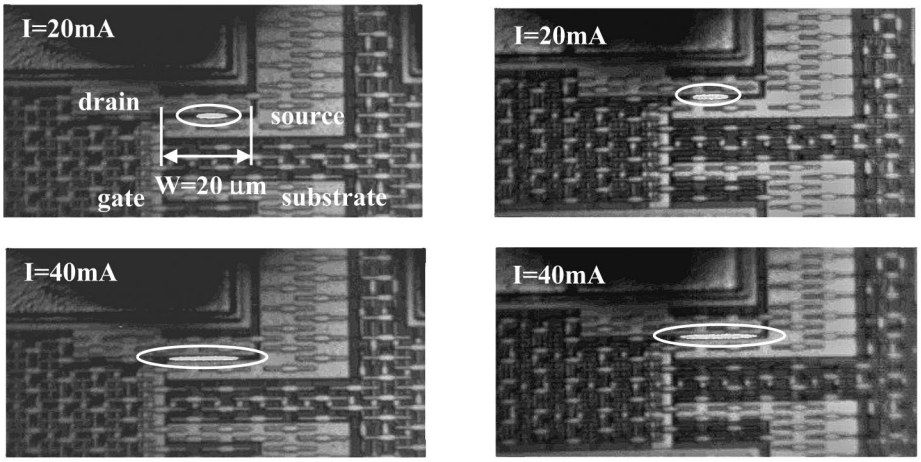
B. EMMI Analysis

In order to visualize the strong nonuniformity of lateral n-p-n bipolar current conduction inferred from the I_{t2} data for the



(a) (b)

Fig. 3. Silicide process dependent I_{t2} with the single-finger width for (a) the 1.5-V and (b) 3.3-V NMOS transistors. The I_{t2} roll-off with W indicates that the failure current essentially remains constant as W is further increased beyond this roll-off point.



(a) (b)

Fig. 4. EMMI images showing spatial extent of lateral current conduction at different current levels for 3.3-V ($W/L_{poly} = 20/0.5$ μm) (a) silicided and (b) nonsilicided single-finger NMOS devices. ($T_p = 300$ ns, $f = 400$ Hz and $T_{\rm exp} = 6$ min).

single-finger transistors, the spatial distribution of ESD current was directly observed using EMMI for the silicided and nonsilicided, 20-μm and 80-μm–wide finger, high-voltage (3.3-V) NMOS transistors. EMMI is a widely used technique for wafer level reliability and yield analysis for semiconductor devices. In general, this analysis is performed by collecting visible and near infrared wavelength (390 $\sim$ 1000 nm) photons emitted under device operation. For NMOS transistors under high electric fields and currents, such as ESD, the radiative intraband transitions of photons are the predominant emission mechanism in EMMI analysis. The generated photons, transmitted through the overlying layers such as dielectrics and metal interconnections, can be detected and this is referred as a front-side light-emission analysis. In this work, the FA-1000 model of EMMI (manufactured by Alpha Innotech Corporation) was utilized and a pulsed bias with duration (T_p) of 300 ns was applied at a frequency of 400 Hz to avoid any thermal failure due to self-heating during the exposure. The emitted photons during each voltage pulse were integrated over the exposure time ($T_{\rm exp}$) of 6 min. In Fig. 4, the observed current distributions are shown at current levels of 20 mA and 40 mA for the 20-μm-wide transistors with silicided and nonsilicided processes. Despite the symmetrical layout of the substrate contacts with respect to the source contacts, at 20 mA of current stress, only a small section of the finger is turned on for both silicided and nonsilicided devices. The turned-on location for each test structure is observed to be inconsistent, which is believed to result from the inhomogeneities in process conditions resulting in statistical random distribution of defects or dopant fluctuations, which causes different device behavior. The turned-on width expands with increased drain current and it is observed that most of the finger width is eventually turned on at a current of 40 mA. The same phenomenon has also been reported in the literature with infrared laser interferometric technique for a 0.35-μm technology [10]. The EMMI results show strong qualitative correlation with the I_{t2} data for high-voltage devices where the normalized failure currents (I_{t2}) are almost constant up to the finger width of 20 μm for both processes [Fig. 3(b)]. For the 80-μm-wide devices as shown in Fig. 5, the initial turned-on location is also randomly placed as observed for the 20-μm-wide device. However, both silicided and nonsilicided devices failed with permanent or partial damage resulting in high drain leakage current, before the bipolar current conduction width extended to the entire 80-μm finger width, at 40 mA and 60 mA, respectively. Throughout the repetitive tests, full triggering of the 80-μm-wide transistors was not observed for either the silicided or the nonsilicided processes. This obser-

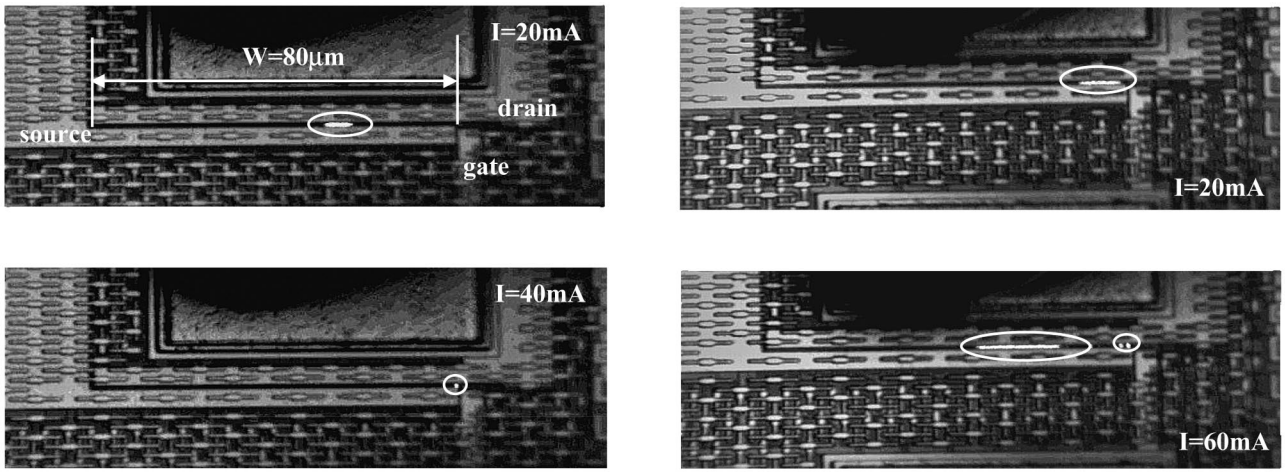


Fig. 5. EMMI images showing spatial extent of lateral current conduction at different current levels for 3.3-V ($W/L_{poly} = 80/0.5$ μm) (a) silicided (b) nonsilicided single-finger NMOS devices ($T_p = 300$ ns, $f = 400$ Hz, $T_{\rm exp} = 6$ min). The small bright spot indicates a failure or a partial failure of the devices. The discontinuity of the conduction region in (b) was not observed in the EMMI images at the lower current level than the 60 mA and the two spots were included within the conduction region.

vation implies that the severe degradation of I_{t2} with increase in (single) finger width results from inhomogeneous bipolar triggering phenomenon and as shown in Fig. 3, the maximum turned-on width ($W_{\rm max}$) of the lateral n-p-n transistor under ESD events can be regarded as the roll-off point in the data of I_{t2} versus W. This qualitatively correlates the results of the EMMI analysis to the TLP measurements. According to the data shown in Fig. 3(a), for low-voltage transistors, the estimated $W_{\rm max}$ value for nonsilicided devices seems to be about 30 μm and for silicided devices, $W_{\rm max}$ seems to be smaller than 5 μm. Moreover, for high-voltage transistors, $W_{\rm max}$ for nonsilicided devices and silicided devices are about 35 μm and 20 μm, respectively. Hence, in practice, designed finger widths greater than $W_{\rm max}$ cause no improvement in I_{t2}. Thus, the obvious way of improving ESD strength is to expand the turned-on width, $W_{\rm max}$ for any given NMOS structures.

III. Physical Modeling of Nonuniform Bipolar Conduction

For better insight into the experimental results, a simple physical model for describing the nonuniform bipolar conduction is proposed. As the experimental results show, the spatial extent of bipolar conduction is determined by the triggered portion of the finger width. The single-finger transistor can be considered as a parallel-connected network of narrow (segmented) n-p-n transistors as shown in Fig. 6. Since each segmented n-p-n transistor has slightly different intrinsic characteristics, which stems from the inherent statistical variations, the location of the triggered segmented n-p-n transistor (or transistors) is expected to be uncertain in this sense. According to the study by Russ *et al.* [7], avalanche multiplication starts at the corner of the drain structure where the field is highest due to the spherical junction curvature and the avalanche current is rather uniformly distributed along the channel width before the snapback of the lateral n-p-n transistor occurs. However, for the devices with STI used in this study, the initially turned-on location along

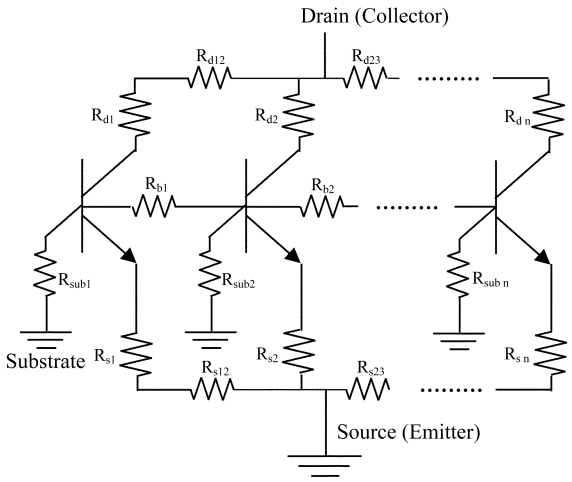
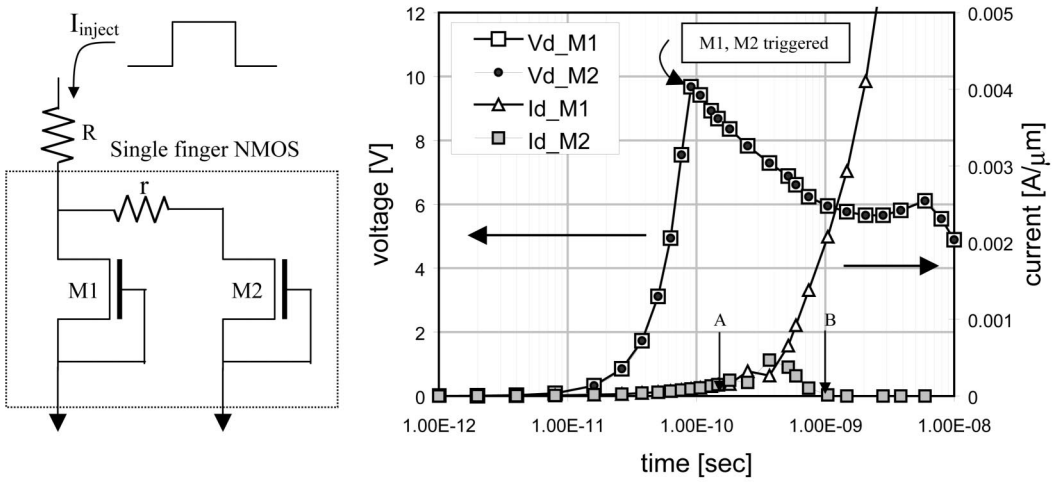
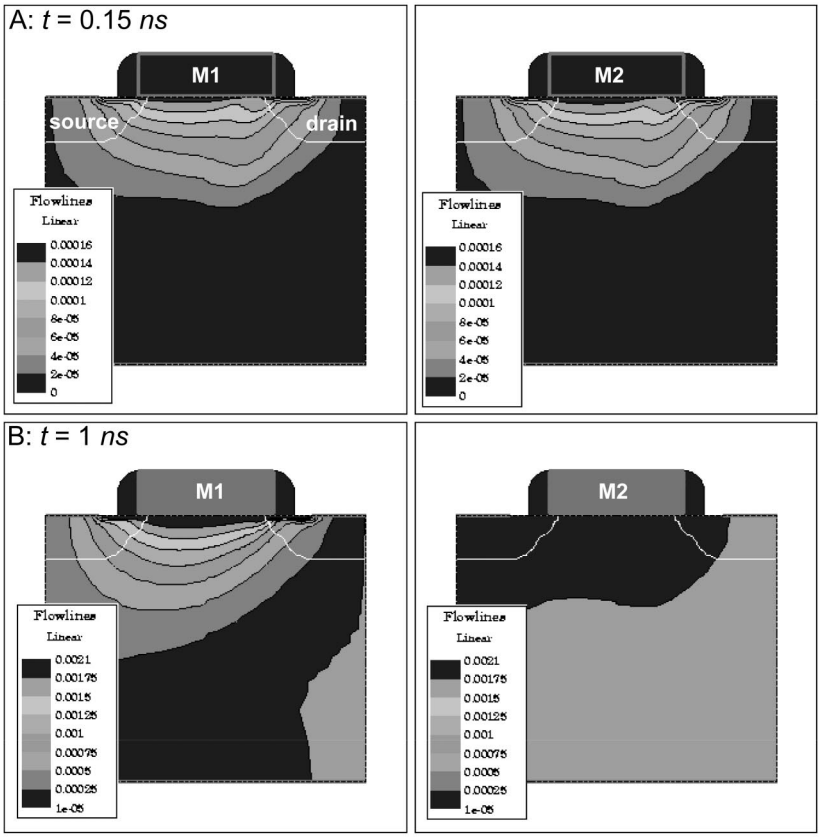


Fig. 6. Schematic of the segmented n-p-n transistors for a gate-grounded single-finger NMOS transistor. Each n-p-n transistor has different intrinsic characteristics due to the statistical variations. R_s, R_d, $R_{\rm sub}$, and R_b denote the parasitic resistance in the source, the drain, the substrate and the intrinsic base, respectively.

the finger width has been observed to be random. Moreover, the avalanche region does not seem to be spread out enough to trigger the entire n-p-n transistor structure for 80-μm-wide single-finger devices. The parasitic bipolar triggering mechanism has been described in [11] in terms of three main device parameters such as the current gain β, the substrate resistance R_{sub} and the multiplication factor M. The substrate hole current (for NMOS) is strongly influenced by the electric field distribution of the drain junction, which depends on the doping profile and drain-engineering of the structures. The resulting effective forward bias to the source-substrate junction by the local substrate potential for each segmented n-p-n transistor is unlikely to be equal due to local variations of the substrate current and substrate resistance. Assuming that a small portion of the finger width (source-substrate junction) is sufficiently ($>\sim 0.8$ V) forward biased because of rather strong impact ionization process,



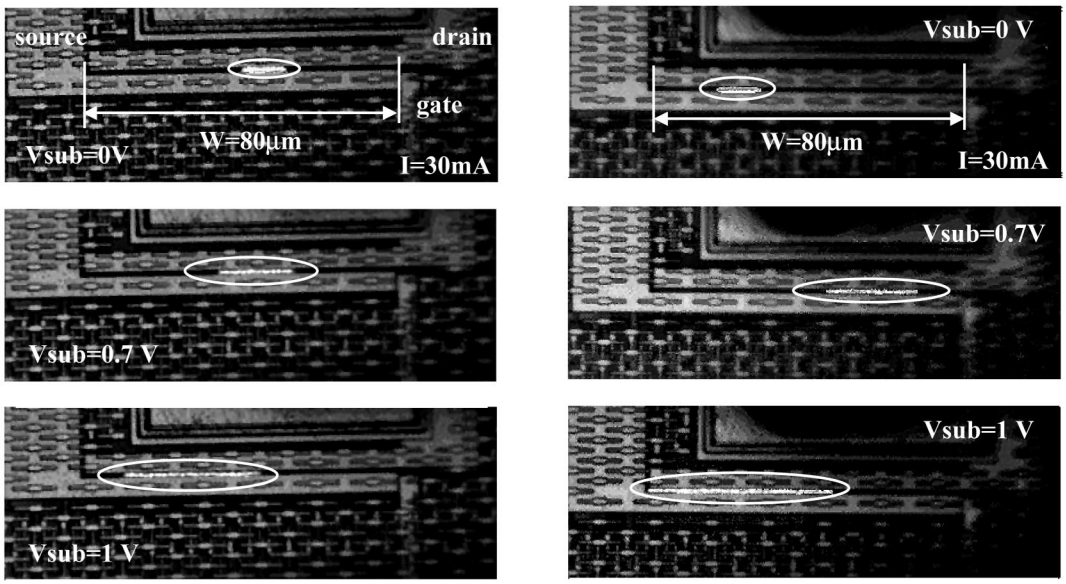
(a) (b)



(c)

Fig. 7. Nonuniform current conduction with a mixed-mode transient simulation for M1 and M2 ($W/L_{poly} = 1/0.5\ \mu$m). (a) The schematic of simulation ($R = 500\ \Omega$ and $r = 0.81\ \Omega$); (b) drain current and voltage with elapsed time for M1 and M2; and (c) the current flowlines at the two different time conditions, A and B.

the segmented transistors within this portion can be immediately triggered while the transistors along the rest of the finger width still remain off and this model agrees well with the EMMI observation that the location of initially turned-on segment is random over the finger width. Once the n-p-n transistors turn on, to maintain the on-state of the transistors, the snapback condition of $\beta \cdot (M-1) > 1$ [12] should be satisfied. However, both β and M are also functions of the injected drain current [11]. Therefore, the location of turned-on segmented n-p-n transistor should be strongly influenced by the drain current. With increase in drain current, the number of turned-on segmented transistors should increase because the maximum current capability for each segmented transistor is limited. This is in agreement with EMMI analysis which showed that the turned-on width spreads out with increase in the injection current. However, even with increased injection current, lateral bipolar currents tend to flow through the portion of the finger width where impact ionization occurs most strongly. The rest of the finger width, where the im-



(a) (b)

Fig. 8. EMMI images that show the spatial extent of lateral current conduction with different $V_{\rm sub}$ for 3.3-V ($W/L_{poly} = 80/0.5$ μm) (a) silicided and (b) nonsilicided single-finger NMOS devices. ($T_p = 300$ ns, $f = 400$ Hz, $T_{\rm exp} = 6$ min).

pact ionization is relatively small, hardly turns on. This happens because the drain voltage drops to the holding voltage V_{sp}, after a part of bipolar transistor structure triggers at V_{t1}. This nonisotropic device behavior along the width is technology dependent and furthermore, even small asymmetric device properties can easily induce this nonuniform current conduction, which can be supported by two-dimensional (2-D) electrothermal transient device simulations.

As shown in Fig. 7(a), two NMOS structures, M1 and M2, are tied together to represent each half of a single-finger transistor. For simplicity, nonisotropic properties of devices are represented by a parasitic resistance (r) between the two drains, which represents variations in drain series resistance. This affects the strength of impact ionization through electric field reduction and also changes the effective series resistance for the two NMOS transistors. The increase in the injection current (with elapsed time) triggers both the transistors, M1 and M2, at $t = 0.09$ ns. After the snapback, the currents for the two transistors increase together up to 0.5 mA/μm as shown in Fig. 7(b). However, as the current increases for the transistor (M2), the increase in voltage drop across the parasitic resistance can reduce the strength of avalanche multiplication of M2. In that case, M2 turns off. This can be observed from the current flowlines shown in Fig. 7(c). The simulation results confirm that the inequality of intrinsic characteristics of each segmented transistor causes asymmetry in current conduction and subsequently results in current localization to become dependent on the injected drain current level.

IV. Bias Dependencies of ESD Robustness

In this section we investigate the substrate and gate-bias dependencies of ESD robustness. As a means of improving the uniformity of ESD current distribution, the substrate triggering and the gate coupling technique have been proposed [3], [5]. These bias conditions have significant impact on the ESD robustness of protection devices and on the effectiveness of the protection design. Accordingly, we investigate the influence of bias conditions on the uniformity of bipolar conduction in single-finger NMOS transistors.

A. Substrate-Bias Effect

For comprehensive understanding of the influence of substrate-bias, EMMI has been performed with an external forward substrate-bias ($V_{\rm sub}$) to the emitter and base (the source and substrate) junction of the NMOS transistor. With increase in substrate-bias, the local substrate potential can be sufficiently raised to trigger the parasitic lateral n-p-n transistor without relying on the self-biasing mode. The influence of the substrate-bias on the spatial current distribution is apparent as shown in Fig. 8. It can be observed that the transistor turned-on width spreads out with increase in $V_{\rm sub}$ for both silicided and nonsilicided devices at the constant drain current of 30 mA.

In fact, the total turned-on width is enlarged by three to four times with $V_{\rm sub} = 1$ V regardless of the silicide process. However, even with the substrate-bias, full n-p-n transistor triggering was not observed for the 80-μm-wide transistors. Nevertheless, EMMI images clearly illustrate that the substrate-bias can increase the effective finger width by extending the lateral bipolar conduction width, which can lead to the improvement of I_{t2} as shown in Fig. 9. It is important to note that while this positive impact of substrate-bias on I_{t2} has been reported before for a 0.35-μm process [3], the physical mechanism responsible for the I_{t2} improvement with $V_{\rm sub}$ has not been adequately explored. It can be observed from Fig. 9 that despite the difference in substrate resistance ($R_{\rm sub}$), the substrate-bias is effective and it also implies that the improvement of the ESD performance can be realized without changes in the process or structure of the devices. As the forward substrate-bias increases, the n-p-n bipolar triggering voltage (V_{t1}) reduces and

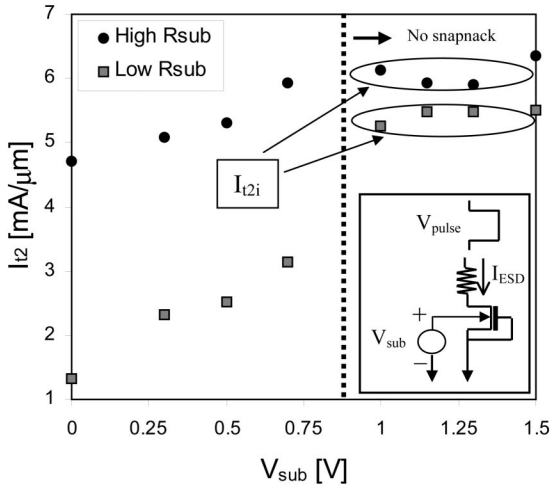


Fig. 9. Second breakdown triggering current I_{t2} with $V_{\rm sub}$ for the 3.3-V silicided devices with two different $R_{\rm sub}$ (6300 Ω-μm and 4800 Ω-μm) where $W/L = 20/0.5$ μm. I_{t2} approaches its intrinsic value, I_{t2i}, as the substrate-bias is increased.

eventually the bipolar turns on without snapback when the effective emitter-base (source-substrate) junction bias reaches $\sim$ 0.8 V. From Fig. 9, at $V_{\rm sub} = 1$ V, it can be inferred that weak avalanche generation is adequate to supply the required substrate current for triggering the lateral n-p-n transistor. For higher $V_{\rm sub}(\geq 1$ V), I_{t2} values tend to saturate. This means that the effective bipolar conduction width gets pinned and the local substrate potential near emitter-base junction cannot be altered by applying additional substrate-bias. The associated I_{t2} at this substrate-bias will be the maximum achievable value for both high and low substrate resistance devices. This current is substrate-bias independent as shown in Fig. 9 and can be called as the intrinsic second breakdown triggering current I_{t2i} for a given technology. In addition, the nonuniform triggering that arises from the inhomogeneity of local substrate potential (along the width of NMOS) can be alleviated by applying a $V_{\rm sub}$ since the effective finger width can be increased by enhanced bipolar current uniformity. For the test structures with various finger widths, I_{t2} values are shown as a function of $V_{\rm sub}$ in Fig. 10. It can be observed that $W_{\rm max}$, where I_{t2} rolls off, increases with substrate-bias and the I_{t2} values for silicided devices approach that of nonsilicided devices with substrate-bias. However, it can also be observed that the values of I_{t2} for $W < W_{\rm max}$ remain almost independent of the substrate-bias within the scatter among the data. Hence, this value of I_{t2} for finger widths less than $W_{\rm max}$ can be thought of as the maximum obtainable I_{t2} (defined as I_{t2i} earlier) under uniform bipolar conduction for a given process technology and ranges from 6 to 7.2 mA/μm for both the processes, which is solely determined by process effects such as silicide/nonsilicide process, gate-to-contact spacing, source/drain engineering and substrate resistance.

Aside from three-dimensional (3-D) effects on ESD current distribution, insight concerning the substrate-bias effect can be attained by 2-D device simulations. In Fig. 11, static high current characteristics of a gate-grounded NMOS transistor with $L_{poly} = 0.5$ μm is simulated for two different substrate-bias

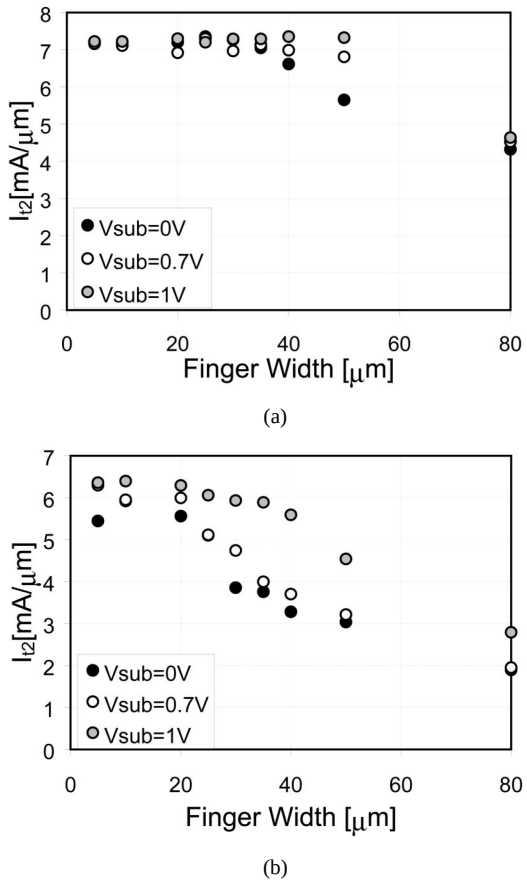


Fig. 10. I_{t2} as a function of transistor width for different substrate-bias for (a) nonsilicided and (b) silicided 3.3-V NMOS transistors.

conditions, $V_{\rm sub} = 0$ V and $V_{\rm sub} = 0.75$ V. The high current I–V curves show the same physical trends as the measured data. At a drain current of 500 μA/μm after n-p-n transistor triggering, the current flowlines are compared for the two substrate-bias conditions, which shows that the current flows more deeply into the substrate with a substrate-bias. The altered local substrate potential changes the snapback triggering voltage, V_{t1} [Fig. 11(a)] and eventually turns on the n-p-n transistor without snapback. However, the I–V curves at high current levels show no significant differences between the two substrate-bias conditions since any 3-D behavior cannot be taken into account. The second snapback at the drain current of ~ 5 mA/μm in the I–V curves is attributed to a rapid increase in the base current resulting from a significant increase in the current component due to thermally generated carriers although the carriers generated by impact ionization decrease slightly with temperature rise. In the case of the self-biasing mode, sufficient substrate current due to impact ionization is required to maintain the forward bias to the emitter and base junction (source and substrate) and the base current is also supplied by impact ionization. Hence, rather strong avalanche multiplication is required for triggering the lateral n-p-n transistor. On the other hand, under adequate external substrate-bias, the lateral n-p-n transistor operates in a normal biasing mode (common emitter). Even in the absence of the drain current, the source-substrate and the drain-substrate junctions are fully turned on. However, since both the parasitic diodes in an NMOS transistor have a relatively long base, most

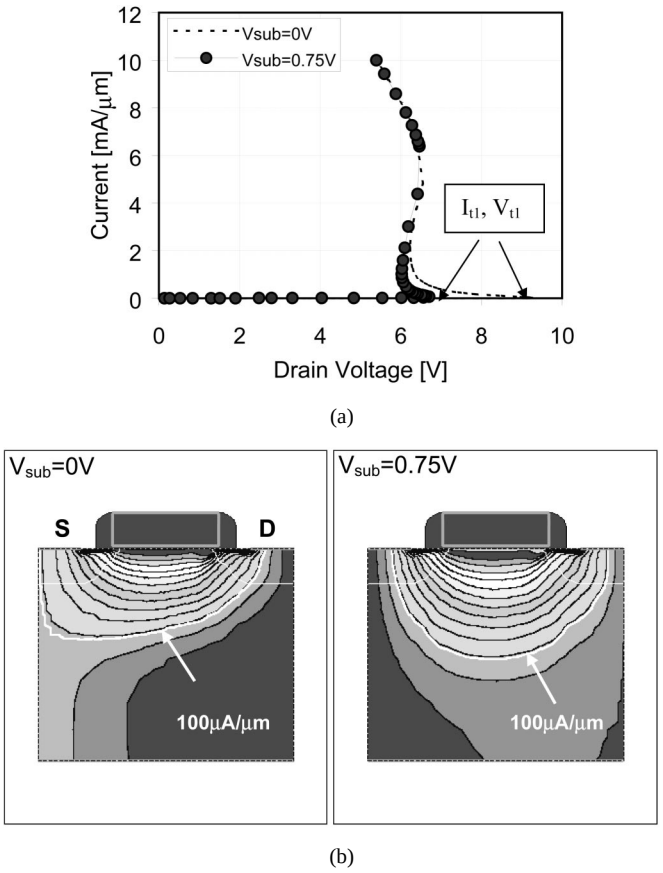


Fig. 11. Static I–V characteristics for the gate-grounded NMOS with $L_{poly} = 0.5\ \mu$m. (a) High-current I–V curve and (b) current flowlines at the drain current (I_D) of 500 μA/μm with $V_{\rm sub} = 0$ V and $V_{\rm sub} = 0.75$ V. The current flows more deeply into the substrate with $V_{\rm sub}$.

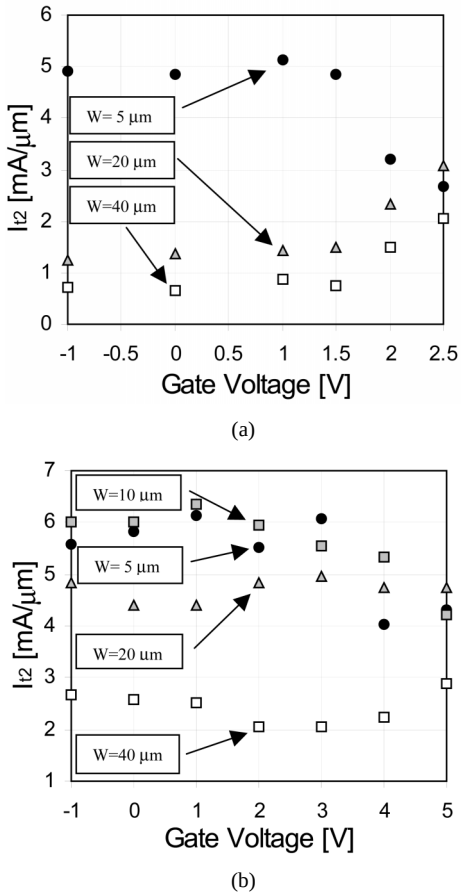
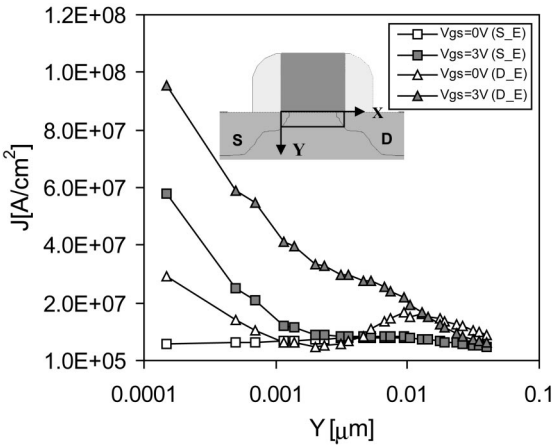


Fig. 12. Second breakdown triggering current (I_{t2}) with gate-bias for the (a) low-voltage NMOS (1.5-V NMOS with $L_{poly} = 0.175\ \mu$m) and (b) high-voltage NMOS transistors (3.3-V NMOS with $L_{poly} = 0.5\ \mu$m).

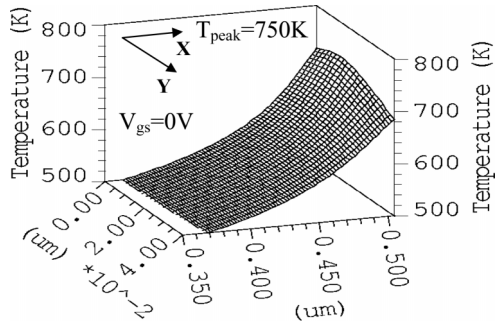
of the injected carriers from the source and drain to the substrate recombine resulting in small diffusion currents. As the drain current and the associated drain bias increase, the drain-substrate junction is eventually reversed biased and thus the lateral n-p-n transistor operates under normal bias conditions. However, the bipolar conduction current increases very slowly with the increase in the drain current (the associated drain bias) until impact ionization is initiated. Before the avalanche multiplication occurs, most of the drain current is supported by electrons injected from the source (emitter), but these injected carriers mainly contribute to the source-substrate diode current rather than to the drain (collector) current due to the low current gain β of the lateral n-p-n transistor arising from the significant electron-hole recombination in the base region. Further increases in the drain current, I_D induces impact ionization and eventually the avalanche multiplication process becomes regenerative as in the self-biasing mode. Since the emitter base junction is already fully turned on, small values of the avalanche-generation current, $I_{\rm gen}$, are sufficient to initiate the regenerative process and accordingly the corresponding threshold value of M should be lower than the value needed in a self-biasing mode. As a result, the drain current flows through the low field area in the drain-substrate junction. Fig. 11(b) clearly shows that relatively wider area of the drain junction is utilized for the same current conduction with $V_{\rm sub}$ because less impact ionization current is required. Therefore, 2-D simulation results suggest that the bipolar conduction could take place over a wider area of the drain and substrate junction under sufficient substrate-bias.

B. Gate-Bias Effect

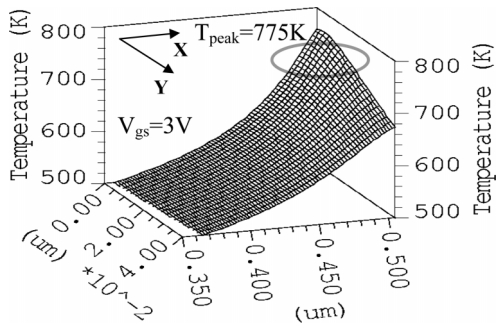
For multifinger NMOS protection devices, the gate coupling technique has been considered to be effective in increasing ESD strength by ensuring uniform triggering of the lateral n-p-n [4], [13], although it is less effective in silicided processes [14]. However, it is also known that excess gate coupling degrades I_{t2} of NMOS devices and thus design techniques have been used to limit the gate coupling [5]. Even with controlled gate coupling to the protection device, ESD failure can occur depending on the gate coupling level, which implies that the ESD strength of the NMOS transistor could be either improved or degraded with gate-bias [15]. However, the physical mechanism for the I_{t2} degradation with gate-bias and its dependence on the finger width for advanced NMOS transistors has not been investigated. In Fig. 12, I_{t2} values are shown with various gate-bias conditions, for low and high-voltage silicided NMOS transistors. It can be observed that the measured I_{t2} of the NMOS transistors is strongly dependent on the applied gate-bias and on the single-finger width as well. For the I_{t2} dependence on the gate-bias, contradictory trends can be clearly seen, depending on the gate finger width of the NMOS transistor. This implies that the gate-bias can result in competing physical mechanisms



(a)



(b)



(c)

Fig. 13. Current density and temperature distribution with different gate-bias conditions by transient simulation with the pulsed current ($I_D = 10$ mA/μm and $t_r = 10$ ns). (a) The current density at the edge of source (S_E) and drain (D_E) extension along y-axis with gate-bias (see the x and y axis in the rectangle underneath the gate). Overall temperature distribution inside the rectangle for (b) $V_{\rm gs} = 0$ V and (c) $V_{\rm gs} = 3$ V.

depending on the finger width for a given structure. It is known from the data in Fig. 3 that the ESD current distribution is uniform within the very narrow finger width devices, such as $W < 5\ \mu$m for the low-voltage (1.5-V) transistors and $W \leq 20\ \mu$m for the high-voltage (3.3-V) transistors. Therefore, it can be inferred that improvement in I_{t2} of the wide finger devices ($W = 20\ \mu$m and $W = 40\ \mu$m) for the low-voltage devices with gate-bias is apparent where the ESD currents are strongly nonuniform. On the other hand, I_{t2} of the narrow finger device ($W = 5\ \mu$m for the 1.5-V NMOS and $W = 5\ \mu$m and 10 μm for 3.3-V NMOS), where ESD current is known to conduct nearly uniformly, is degraded with gate-bias. As is well known, boosting of the substrate current with gate-bias can alleviate the current

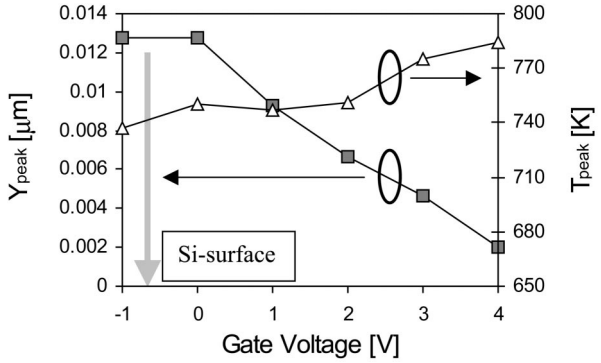
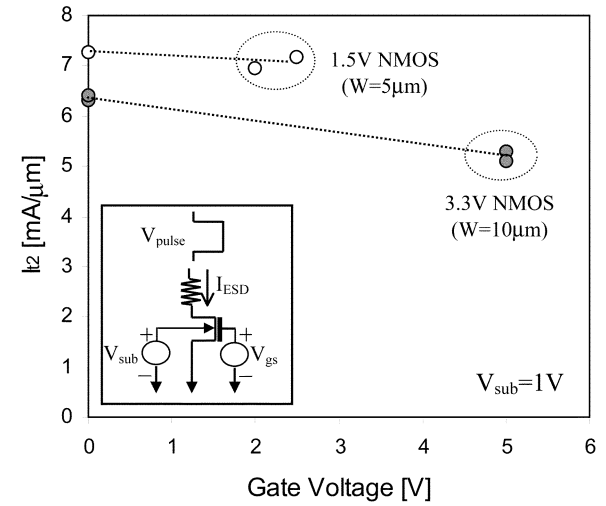
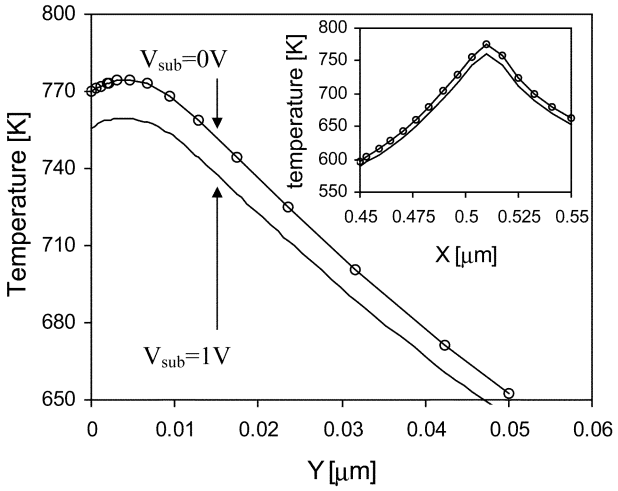


Fig. 14. Location of peak temperature ($Y_{\rm peak}$) and the peak temperature ($T_{\rm peak}$) with gate-bias. With increase in gate-bias, $Y_{\rm peak}$ moves closer to the surface and the peak temperature and $T_{\rm peak}$ also moderately increases.

localization problem by ensuring uniform n-p-n triggering. This mechanism seems to work for the wide finger 1.5-V NMOS devices [Fig. 12(a)] with considerable improvement of I_{t2}, while the improvement in I_{t2} for the wide finger 3.3-V devices is less apparent [Fig. 12(b)]. However, the severe reduction in I_{t2} with gate-bias for the narrow transistors seems to be independent of the turn-on efficiency of the parasitic lateral n-p-n structure. To identify the root cause of this degradation with gate-bias, transient electro-thermal simulations have been employed for the NMOS with $L_{poly} = 0.175\ \mu$m. The simulations in Fig. 13(a) show that the current density within the source/drain extension junction depth is strongly modulated by gate-bias. This implies that the distribution of the local temperature rise in the drain extension and the channel area [indicated by the rectangle inside the NMOS in Fig. 13(a)] can also be influenced by the applied gate-bias. At a drain current of 10 mA/μm (at $t = 10$ ns), the local temperature distribution in the box are shown in Fig. 13(b) and (c). The simulation results show that the distribution of the local temperature near the channel area increases as gate-bias increases. In addition, the simulation shows that the location of the peak temperature resides in the drain extension and it moves closer to the surface with gate-bias as shown in Fig. 14. Hence, this heating effect induced by the gate-bias can lead to I_{t2} degradation in devices where the lateral ESD currents flow uniformly. Also, for negative gate-bias, the location of the peak temperature does not change at all. These simulation results suggest that I_{t2} remains the same with the negative gate-bias. This interpretation agrees well with the measured data in Fig. 12. As the gate-bias increases, the surface heating becomes stronger since the location of the peak temperature approaches the Si/SiO_2 interface. This means that more heat can be accumulated near the surface with gate-bias and the device tends to be more vulnerable to thermal failures at the surface. To verify this heating effect, I_{t2} was also measured with both the gate and the substrate-bias as shown in Fig. 15. It was observed that the reduction in I_{t2} nearly disappeared with substrate-bias, since the lateral ESD current conduction occurs more deeply in the silicon substrate with substrate-bias, leading to reduced heating near the surface. The simulation results also support the experimental observation that the substrate-bias can somewhat reduce the surface heating. Hence, it can be concluded that gate-bias induced heating effect pri-

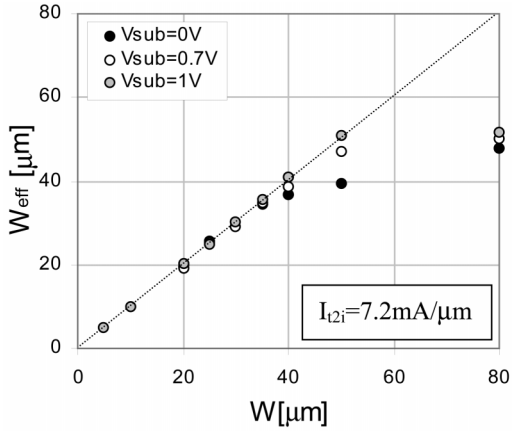


(a)

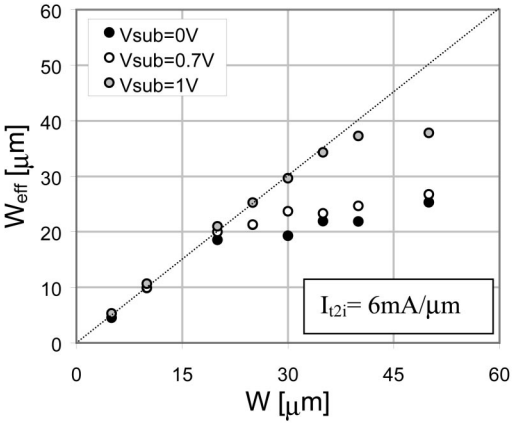


(b)

Fig. 15. Effect of substrate-bias on the I_{t2} degradation with gate-bias. (a) I_{t2} with V_{gs} when $V_{sub} = 1$ V for the low and high-voltage devices (b) simulated temperature rise for the low-voltage device with $V_{gs} = 3$ V where $T_{peak} = 775$ K for $V_{sub} = 0$ V and $T_{peak} = 760$ K for $V_{sub} = 1$ V.



(a)



(b)

Fig. 16. Effective finger widths (W_{eff}) versus designed finger widths (W) for 3.3-V NMOS transistors with different substrate-bias (a) nonsilicided devices and (b) silicided devices.

marily accounts for the reduction in I_{t2} for devices with uniform lateral ESD current conduction.

V. Implications for the Design of ESD Protection

As is well known, the ESD strength of silicided technology is lower than that of nonsilicided devices due to either the reduction in emitter efficiency [3] or early current localization associated with the reduced series resistance [16], which has also been verified in this work. EMMI analysis shows that different bipolar turned-on widths can be obtained depending on the substrate and gate-bias conditions at a given ESD current level. Therefore, these facts indicate that wider current conduction is associated with a higher ESD failure threshold for a given protection structure. For advanced silicided NMOS transistors, I_{t2} drops off rapidly beyond W_{max}. This obviously places a severe restriction on determining the useful width of a single-finger for multifinger structures used in ESD protection. From a practical design point of view, the minimum value for W_{max} should be at least 30 μm with a minimum I_{t2} (at $T_p = 200$ ns) of $\sim$ 4 mA/μm. This will ensure that $>$ 8 V/μm for HBM is available for the design of multifinger protection devices. Based on the value of I_{t2i} for each technology, the effective finger width W_{eff}, for bipolar conduction can be determined from the simple relation $W_{eff} = (I_{t2}/I_{t2i}) \cdot W$ as shown in Fig. 16. At $V_{sub} = 1$ V, W_{max} is 50 μm and 35 μm for the nonsilicided and silicided technologies respectively, showing significant improvement over the W_{max} values for the zero substrate-bias case. Therefore, the use of substrate-bias can extend ESD design capabilities beyond present design and technology limits and one potential application has recently been implemented [17]. In addition to the impact of substrate-bias, the gate-bias dependence has been quantified with the experimental results for 3.3-V NMOS transistor in Fig. 17. Depending on the finger width, two competing trends of the gate-bias effect are clearly observed. Combining the overall results and considering both the impact of the gate-bias and the substrate-bias, a design window for advanced protection devices can be established as shown in Fig. 18. It should be noted that for substrate trigger protection [8], [17], the I_{t2} roll-off with gate-bias is less important. Since the substrate-bias can compensate for the adverse effects of gate-bias, protection devices can be designed with either the gate grounded or gate coupled configurations, as long

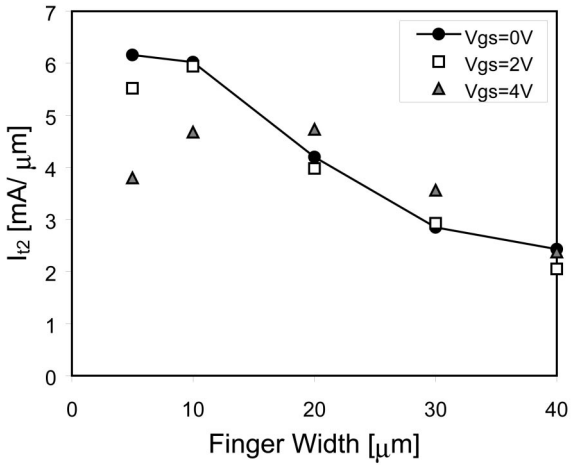


Fig. 17. I_{t2} of the high-voltage (3.3-V) NMOS transistor with finger widths for the various gate voltages. Two competing trends are clearly shown.

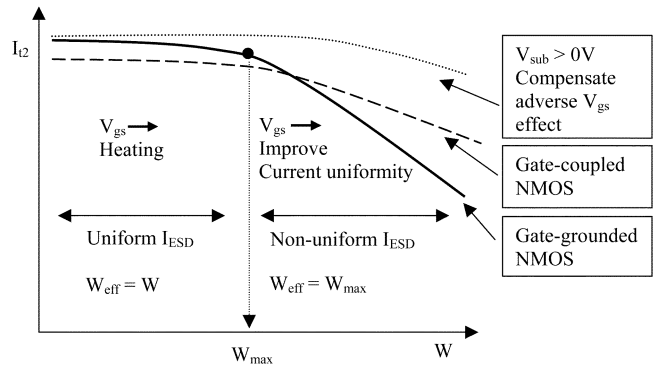


Fig. 18. Design window for optimizing the performance of deep submicron ESD protection.

as the substrate-bias is efficiently supplied for multifinger n-p-n structures. On the other hand, for the design of gate coupled ESD protection devices without substrate-bias, the gate potential of the protection device should be controlled with reasonable values of coupling resistance and capacitance to maintain its value below the level above which I_{t2} begins to roll-off with the gate-bias.

VI. Conclusion

In conclusion, we have shown that nonuniform bipolar conduction phenomenon in advanced single-finger NMOS transistors results in severe reduction in ESD protection strength depending on the device finger width. Also, the impact of substrate and gate-bias conditions on this phenomenon has been explored both experimentally and using 2-D device simulations. The results provide improved understanding of ESD behavior and new physical insight into the substrate-bias and the gate-bias effects involved in advanced ESD protection devices. It is shown that ESD performance can be improved with the substrate-bias by enlarging the effective turned-on device width. Additionally, the concept of an intrinsic second breakdown triggering current (I_{t2i}) is introduced, which is substrate-bias independent and represents the maximum achievable ESD failure strength for a given technology. It is also shown that gate-bias induced heating near the drain extension region close to the Si/SiO_2 surface is the primary cause of the degradation of ESD performance for the devices with uniform bipolar conduction. Moreover, it is established that substrate-biasing can help eliminate the negative impact of the gate-bias effect. Results from this work can be used to construct suitable design windows for efficient and robust ESD protection design to overcome ESD failures in advanced deep-submicron CMOS technologies.

References

[1] D. Scott, J. Hall, and G. Giles, "A lumped element model for simulation of ESD failures in silicided devices," in *Proc. EOS/ESD Symp.*, 1986, pp. 41–47.

[2] T. Polgreen and A. Chatterjee, "Improving the ESD failure threshold of silicided n-MOS output transistors by ensuring uniform current flow," *IEEE Trans. Electron Devices*, vol. 39, pp. 379–388, Feb. 1992.

[3] A. Amerasekera, C. Duvvury, V. Reddy, and M. Rodder, "Substrate triggering and salicide effects on ESD performance and protection circuit design in deep submicron CMOS processes," in *IEDM Tech. Dig.*, 1995, pp. 547–550.

[4] C. Duvvury and C. Diaz, "Dynamic gate coupled NMOS for efficient output ESD protection," in *Proc. IRPS*, 1992, pp. 141–150.

[5] J. Z. Chen, A. Amerasekera, and C. Duvvury, "Design methodology for optimizing gate driven ESD protection circuits in submicron CMOS processes," in *Proc. EOS/ESD Symp.*, 1997, pp. 230–239.

[6] P. Salome, C. Leroux, J. P. Chante, P. Crevel, and G. Reimbold, "Study of a 3D phenomenon during ESD stresses in deep submicron CMOS technologies using photon emission tool," in *Proc. IRPS*, 1997, pp. 325–332.

[7] C. Russ, K. Bock, M. Rasras, and I. D. Wolf, "Non-uniform triggering of gg-nMOST investigated by combined emission microscopy and transmission line pulsing," in *Proc. EOS/ESD Symp.*, 1998, pp. 177–186.

[8] K.-H. Oh, C. Duvvury, C. Salling, K. Banerjee, and R. W. Dutton, "Non-uniform bipolar conduction in single finger NMOS transistors and implications for deep submicron ESD design," in *Proc. IRPS*, 2001, pp. 226–234.

[9] T. Maloney and N. Khurana, "Transmission line pulsing techniques for circuit modeling of ESD phenomena," in *Proc. EOS/ESD Symp.*, 1985, pp. 49–54.

[10] M. Litzenberger, K. Esmark, D. Pogany, C. Furbock, H. Gossner, E. Gornik, and W. Fichtner, "Study of triggering inhomogeneities in gg-nMOS protection devices via thermal mapping using backside laser interferometry," *Microelectron. Reliab.*, vol. 40, pp. 1359–1364, 2000.

[11] V. Gupta, A. Amerasekera, S. Ramaswamy, and A. Tsao, "ESD-related process effects in mixed-voltage sub-0.5 μm technologies," in *Proc. EOS/ESD Symp.*, 1998, pp. 161–169.

[12] R. W. Dutton, "Bipolar transistor modeling of avalanche generation for computer circuit simulation," *IEEE Trans. Electron Devices*, vol. ED-22, pp. 334–338, 1975.

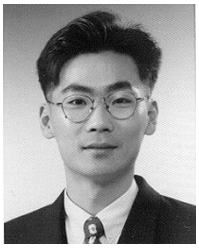
[13] C. Duvvury, C. Diaz, and T. Haddock, "Achieving uniform nMOS device power distribution for sub-micron ESD reliability," in *IEDM Tech. Dig.*, 1992, pp. 131–134.

[14] S. Ramaswamy, C. Duvvury, and S.-M. Kang, "EOS/ESD reliability of deep sub-micron NMOS protection devices," in *Proc. IRPS*, 1995, pp. 284–291.

[15] K.-H. Oh, C. Duvvury, K. Banerjee, and R. W. Dutton, "Gate bias induced heating effect and implications for the design of deep submicron ESD protection," in *IEDM Tech. Dig.*, 2001, pp. 315–318.

[16] G. Notermans, A. Heringa, M. V. Dort, S. Jansen, and F. Kuper, "The effect of silicide on ESD performance," in *Proc. IRPS*, 1999, pp. 154–158.

[17] C. Duvvury, S. Ramaswamy, A. Amerasekera, and R. A. Cline, "Substrate pump NMOS for ESD protection application," in *Proc. EOS/ESD Symp.*, 2000, pp. 7–17.



Kwang-Hoon Oh (S'92) received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, Korea, in 1990 and 1992, respectively. He is currently pursuing the Ph.D. degree at Stanford University, Stanford, CA.

From 1992 to 1997, he was with Samsung Electronics, Puchon, Korea, where he was engaged in the design and development of power MOSFETs and IGBTs. During 2000–2001, he held summer research positions at Texas Instruments, Inc., Dallas, TX, where he focused on the modeling of ESD reliability for advanced CMOS technologies. His research interests are in the area of device simulation, characterization, and electrothermal and reliability modeling for advanced deep submicron CMOS technologies with applications to IC circuits.

Mr. Oh is a member of the IEEE Electron Devices Society.

Charvaka Duvvury (SM'01) received the Ph.D. degree in engineering science from the University of Toledo, Toledo, OH.

After working as a Postdoctoral Fellow in physics at the University of Alberta, Alberta, ON, Canada, he joined Texas Instruments, Dallas, TX, in 1977. He initially worked in the Houston DRAM Group as a Design/Product Engineer for 4K/16K DRAMS. He then was part of the first 256K CMOS DRAM design and the Advanced Development Group that worked on the 1 Meg DRAM with specific contributions in DRAM circuit design, transistor modeling, and reliability studies. He joined the Semiconductor Process and Device Center, Dallas, in 1988, where his work was on the transistor modeling of CMOS/BiCMOS technologies and development of ESD protection for high voltage designs and submicron CMOS technologies. He was elected Senior Member of Technical Staff in 1990, Distinguished Member of Technical Staff in 1997, and Texas Instruments Fellow, also in 1997. His current work is on ESD development for deep submicron CMOS technologies. He has published over 65 papers in technical journals and conferences and holds 25 patents with several pending. He has coauthored books on hot carriers (New York: Van Nostrand Reinhold, 1992), modeling of electrical overstress (Boston, MA: Kluwer, 1994), and ESD reliability phenomena and protection design (New York: Wiley, 1995).

Dr. Duvvury is a recipient of the Outstanding Contributions Award from the EOS/ESD Symposium (1990), Outstanding Mentor Award from the SRC (1994), several Best Paper Awards from the EOS/ESD Symposium, and Outstanding Paper Award from the International Reliability Physics Symposium. He has been very active in the EOS/ESD Symposium, where he was the Technical Program Chairman of the 1992 Symposium and was the General Chairman of the 1994 ESD Symposium. He is currently a member of the ESD Association Board of Directors, promoting university education and research in ESD. He is also a member of Eta Kappa Nu and Sigma Xi.

Kaustav Banerjee (M'99) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1999.

He was with Stanford University, Stanford, CA, from 1999 to 2002 as a Research Associate at the Center for Integrated Systems. In July 2002, he joined the Faculty of the Department of Electrical and Computer Engineering, University of California, Santa Barbara, as an Assistant Professor. His research interests include nanometer scale circuit effects and their implications for high-performance/low-power VLSI and mixed-signal designs and their design automation methods. He is also interested in some exploratory interconnect and circuit architectures such as 3-D ICs, integrated optoelectronics, and nanotechnologies such as single electron transistors. He co-advises several doctoral students at Stanford University, University of Southern California, Los Angeles, and the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. From February 2002 to August 2002, he was a Visiting Professor at the Circuit Research Labs of Intel, Hillsboro, OR. In the past, he has also held summer/visiting positions at Texas Instruments, Inc., Dallas, TX, and EPFL-Switzerland, and has consulted for several EDA companies in the Silicon Valley. He has authored or coauthored over 70 technical papers in archival journals and refereed international conferences and has presented numerous invited talks and tutorials.

Dr. Banerjee served as Technical Program Chair of the 2002 IEEE International Symposium on Quality Electronic Design (ISQED '02), and is the Conference Vice-Chair of ISQED '03. He has also served on the technical program committees of the ACM International Symposium on Physical Design, the EOS/ESD Symposium, and the IEEE International Reliability Physics Symposium. He is the recipient of a Best Paper Award at the 2001 Design Automation Conference.

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 currently Professor of electrical engineering, Stanford University, Stanford, CA, and Director of Research at 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 processes, 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 doctoral students.

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