CHISEL Programming Operation of Scaled NOR Flash EEPROMs—Effect of Voltage Scaling, Device Scaling and Technological Parameters

Nihar R. Mohapatra, *Student Member, IEEE*, Deleep R. Nair, *Student Member, IEEE*, S. Mahapatra, V. Ramgopal Rao, *Senior Member, IEEE*, S. Shukuri, and Jeff D. Bude

***Abstract*—The impact of programming biases, device scaling and variation of technological parameters on channel initiated secondary electron (CHISEL) programming performance of scaled NOR Flash electrically erasable programmable read-only memories (EEPROMs) is studied in detail. It is shown that CHISEL operation offers faster programming for all bias conditions and remains highly efficient at lower biases compared to conventional channel hot electron (CHE) operation. The physical mechanism responsible for this behavior is explained using full band Monte Carlo simulations. CHISEL programming efficiency is shown to degrade with device scaling, and various technological parameter optimization schemes required for its improvement are explored. The resulting increase in drain disturbs is also studied and the impact of technological parameter optimization on the programming performance versus drain disturb tradeoff is analyzed. It is shown that by judicious choice of technological parameters the advantage of CHISEL programming can be maintained for deeply scaled electrically erasable programmable read-only memory (EEPROM) cells.**

***Index Terms*—channel hot electron (CHE), channel initiated secondary electron (CHISEL), device scaling, Flash electrically erasable programmable read-only memories (EEPROMs), hot carriers, monte carlo simulation, programming efficiency.**

I. Introduction

CHANNEL initiated secondary electron (CHISEL) injection has been shown to be an excellent programming mechanism for NOR Flash electrically erasable programmable read-only memories (EEPROMs) [1]–[5]. It relies on impact ionization feedback, is activated by the application of a negative substrate bias (V_B), and provides high-energy electrons that get injected into the floating gate (FG) over a spatially broader area in the channel [1]–[9], as schematically shown in Fig. 1. When compared against the conventional channel hot electron (CHE) process [10], CHISEL injection provides lower voltage and lower power operation for equivalent programming time (T_P) and faster T_P for equivalent programming power. CHISEL injection also offers self-convergent programming leading to excellent threshold voltage (V_T) control for large arrays with minimum use of program verification and a unique recovery procedure for over erased cells [1]–[5], [9], [11], [12] not available under CHE programming [13].

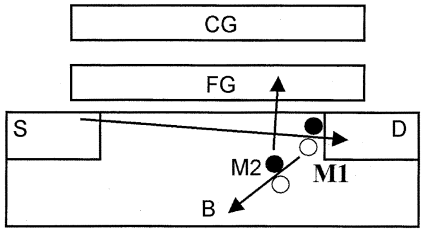


Fig. 1. Schematic of CHISEL injection mechanism. Channel electrons, heated by lateral electric field create primary electron-hole pairs (EHP) by impact ionization (M1). Primary holes flow to the substrate and in the presence of high transverse electric field (due to negative substrate bias) create secondary EHP by impact ionization (M2). The secondary electrons move toward the interface and those having energy greater than 3.1 eV get injected into the FG [1]–[9].

From the reliability perspective, CHISEL programming has shown good programmed and erased V_T endurance up to 10^5 program/erase cycles [14], [15]. Program/erase cycling and data retention (after cycling) results obtained from large 32 Mbit arrays showed tight V_T control and over 10 years of charge retention [14]. It has been clearly demonstrated that CHISEL programming is free from anomalous bit failure through increased disturbs, window closure or charge loss [14].

However to the best of our knowledge, to date there has been no systematic study on CHISEL programming performance of scaled Flash cells. The impact of programming biases and technological parameters on CHISEL programming efficiency of scaled Flash cells remained unexplored. All the results reported so far are either from cells having larger FG length $(L_{\rm FG})$ [7], [8], [16], or from a particular technology and measured at restricted bias conditions [5]. Moreover, there have been very few reports on the scaling properties of NOR Flash programming using CHISEL injection, which has been shown to degrade at lower $L_{\rm FG}$ values [5], [16], [17]. Although it has been suggested that CHISEL injection efficiency of lower $L_{\rm FG}$ cells could be restored by proper optimization of technological parameters, the reported results were either from large $L_{\rm FG}$ cells [16], [17] or from equivalent MOS transistors measured at a fixed gate and drain bias [5]. The latter approach seriously questions the applicability of these results on real Flash EEPROMs, where the FG voltage $(V_{\rm FG})$ changes from a high to low value during the programming operation.

Manuscript received March 11, 2003; revised July 15, 2003. The review of this paper was arranged by C.-Y. Lu.

N. R. Mohapatra, D. R. Nair, S. Mahapatra and V. R. Rao are with the Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai 400 076, India (e-mail: souvik@ee.iitb.ac.in).

S. Shukuri is with the Semiconductor & Integrated Circuits Group, Hitachi, Tokyo, Japan.

J. D. Bude is with Agere Systems, Allentown, PA, USA.

Digital Object Identifier 10.1109/TED.2003.817275

The scope of this paper is twofold. First, we show the impact of control gate ($V_{\rm CG}$) and drain (V_D) biases on CHISEL programming efficiency of *short* $L_{\rm FG}$ *cells*. Second, we study the impact of $L_{\rm FG}$ scaling and variation in technological parameters [channel doping, source/drain (S/D) junction depth and halo] on CHISEL programming efficiency of *actual Flash cells*. We show that CHISEL programming always results in faster T_P under any combination of $V_{\rm CG}$ and V_D, while identical T_P can be achieved by using lower $V_{\rm CG}$ and/or V_D when compared against conventional CHE operation. The possible physical mechanisms responsible for this behavior are identified using full band Monte-Carlo device simulations. It is shown that CHISEL programming efficiency degrades at lower $L_{\rm FG}$, consistent with earlier reports [5], [16], [17]. Based on the results of [5], the transverse electric field ($E_{\rm TRAN}$) and CHISEL programming efficiency are increased by the following ways:

1) heavier channel doping
2) shallower S/D junction and
3) halo implants.

However, an increase in $E_{\rm TRAN}$ also increases drain disturbs [18]. The tradeoff between increase in CHISEL programming efficiency and decrease in program/disturb margin is investigated in detail. We show that by judicious choice of technological parameters the advantage of CHISEL programming can be maintained as $L_{\rm FG}$ is scaled.

II. Experimental

The devices used in this work were designed on a state-of-the-art 0.18 μm triple well process featuring advanced modules such as shallow trench isolation (STI) and self-aligned S/D contacts leading to highly scaled cell area of about 0.45 μm^2. Measurements were performed on isolated, fully scaled Flash cells having $L_{\rm FG}$ of 0.17 μm through 0.34 μm, width (W) of 0.3 μm, tunnel oxide ($T_{\rm OX}$) and oxide–nitride–oxide interpoly-dielectric (ONO IPD) thickness of 12 and 20 nm, respectively and a gate coupling (α_G) of about 0.6. Four types of cell doping schemes were used to study the impact of technological parameters; A) low channel doping, no halo implant and high S/D junction depth (90 nm); B) high channel doping, no halo implant and high S/D junction depth; C) high channel doping, no halo implant and low S/D junction depth (65 nm) and; D) low channel doping, halo implant and high S/D junction depth. Fig. 2 shows the V_T as a function of $L_{\rm FG}$ for all the above doping schemes, measured on identical FG contacted devices having no IPD and control gate. Table I shows the drain current and injection efficiency of type A through D cells under CHE and CHISEL operation. Note that all cells show lower drain current and higher injection efficiency under CHISEL programming operation.

The bias dependent studies (impact of $V_{\rm CG}$ and V_D) were performed on cells having $L_{\rm FG}$ of 0.26 μm. The $L_{\rm FG}$-dependent studies (impact of scaling and technological parameters) were performed on type B cells with $V_{\rm CG}$ and V_D of 8 and 3.7 V, respectively. CHISEL programming was performed at $V_B = -2$ V. For comparison (programming efficiency, impact of programming biases and $L_{\rm FG}$ scaling), CHE programming was performed at $V_B = 0$ V. The source was always grounded during CHE and CHISEL programming and erase. Uniform channel erase was done at $V_{\rm CG} = -20$ V. The efficiency optimization of scaled cells (by variation of technological parameters) was done for CHISEL operation only. Efficiency optimization for CHE programming is beyond the scope of the present paper.

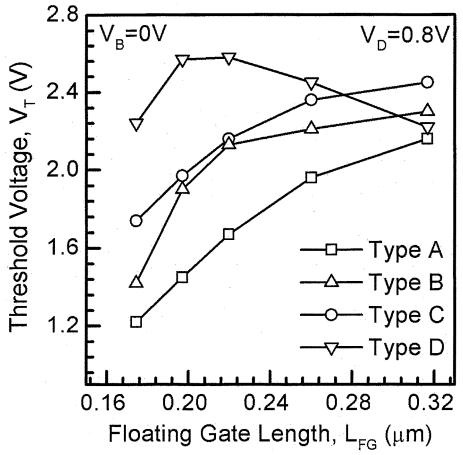
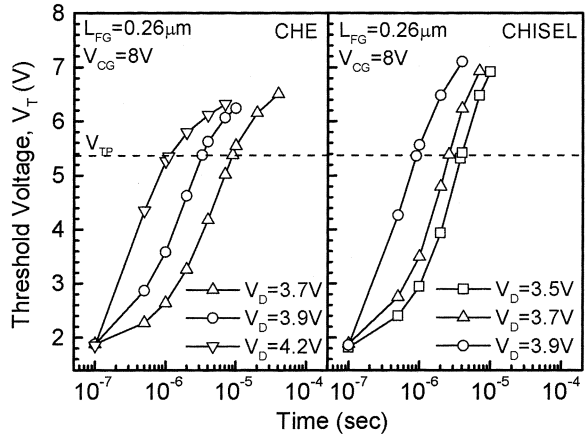


Fig. 2. Threshold voltage as a function of FG length for different doping optimization schemes (Type A through D). Measurements were performed on equivalent floating-gate contacted devices. V_T was defined as V_G required to have $I_D = 5$ μA at $V_D = 0.8$ V.

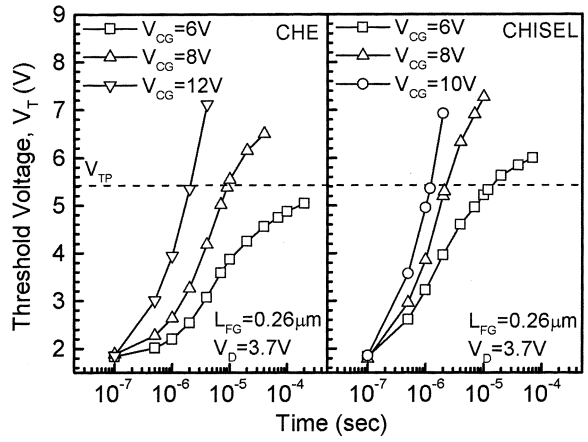
TABLE I
THE DRAIN CURRENT AND INJECTION EFFICIENCY OF 0.26 μm TYPE A, B, C, AND D FLASH CELLS UNDER CHE AND CHISEL OPERATION. THE DATA IS SHOWN FOR $V_{\rm FG}$ AND V_D OF 5 AND 3.7 V, RESPECTIVELY. THE DRAIN AND GATE CURRENTS ARE MEASURED ON EQUIVALENT FLOATING GATE-CONTACTED DEVICES

Type of the cell	L_{FG} (μm)	I_D (μA) (V_D=3.7V, V_{FG}=5V)		I_G/I_D (V_D=3.7V, V_{FG}=5V)	
		CHE	CHISEL	CHE	CHISEL
A	0.26	135.1	98.4	1.75×10^{-7}	7.815×10^{-7}
B	0.26	123	85.1	2.95×10^{-7}	1.603×10^{-6}
C	0.26	115.5	78.3	2.3×10^{-7}	2.35×10^{-6}
D	0.26	107.6	71.5	5.2×10^{-7}	8.713×10^{-6}

Simulations were performed on $L_{\rm FG} = 0.26$ μm FG contacted device having structure and doping profile identical to the measured cell. The structure and doping was obtained from a well-calibrated process simulation, whose output was used as the basis for device simulation. After obtaining the electrostatic potential distribution of the device from drift-diffusion simulation, hot-carrier simulations were performed using SMC, a physics based full band Monte Carlo simulator [9]. The simulated drain (I_D), substrate (I_B) and gate (I_G) currents are in good agreement with the measured data.

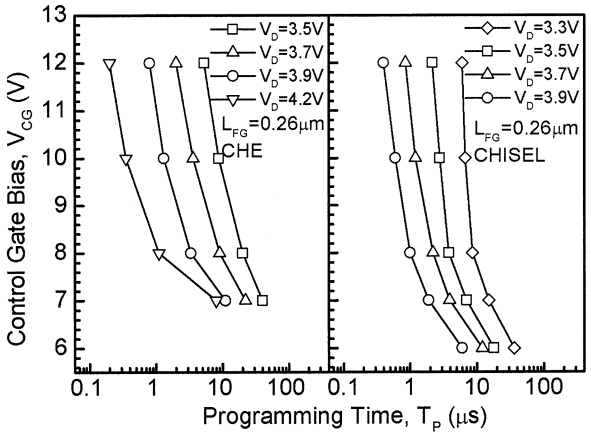


(a)

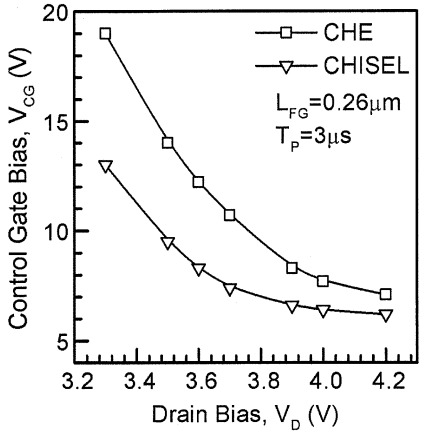


(b)

Fig. 3. Threshold voltage transients of a $L_{\rm FG} = 0.26\ \mu$m cell under CHE and CHISEL programming operation as a function of (a) V_D ($V_{\rm CG}$ fixed at 3.7 V) and (b) $V_{\rm CG}$ (V_D fixed at 8 V). V_T was defined as $V_{\rm CG}$ required to have $I_D = 5\ \mu$A at $V_D = 0.8$ V.



(a)



(b)

Fig. 4. (a) Control gate bias required to obtain a given programming time under CHE and CHISEL operation. (b) Control gate bias required for a given drain bias to achieve a programming time of 3 μs under CHE and CHISEL operation. Measurements were done on a $L_{\rm FG} = 0.26\ \mu$m Flash cell for different sets of V_D values. T_P was calculated for a 3.5 V program V_T shift.

III. Results and Discussion

A. Effect of Programming Biases

All the results reported in this section are from type $BL_{\rm FG} = 0.26\ \mu$m cells. The behavior of other type of cells (A, C and D) are similar in nature.

Fig. 3(a) and (b) show the CHE and CHISEL programming transients under different V_D ($V_{\rm CG}$ constant) and $V_{\rm CG}$ (V_D constant) respectively. For any combination of $V_{\rm CG}$ and V_D, CHISEL always shows faster programming compared to CHE operation. This observation is consistent with previous results [1]–[9], but verified here on a smaller $L_{\rm FG}$ cell and over a wide range of bias choices. For a given V_D, the programming transients measured under low to moderate $V_{\rm CG}$ slow down (saturate) at longer times. Though present under CHISEL operation (at relatively lower $V_{\rm CG}$), this effect is more pronounced under CHE operation unless very high $V_{\rm CG}$ is used. Therefore, CHE programming requires higher V_D and/or $V_{\rm CG}$ to attain a given V_T shift in a given time compared to CHISEL programming. Furthermore, at constant $V_{\rm CG}$, CHE programming transient is faster initially when higher V_D is used but it saturates at longer times. In general, the programming transients start to slow down when a particular program V_T is reached and the saturation sets earlier in time when faster programming is achieved by applying higher V_D.On the other hand, for CHISEL operation this saturation takes place at higher V_T. Therefore, using CHISEL, higher programmed V_T level can be achieved at a much faster T_P compared to CHE programming. The physical phenomenon behind V_T saturation is explained later in this subsection.

Fig. 4(a) shows $V_{\rm CG}$ required to achieve a given T_P (V_T shift of 3.5 V) at different V_D under CHE and CHISEL operation. It is clearly evident that compared to CHE, CHISEL programming offers faster T_P at any given combination of $V_{\rm CG}$ and V_D, while a given T_P is realized with lower $V_{\rm CG}$ and/or V_D. To get a better understanding of CHISEL low-voltage operation, Fig. 4(b) shows $V_{\rm CG}$ required for a given V_D to achieve a T_P of 3 μs under CHE and CHISEL operation. Compared to CHE at any given V_D, CHISEL requires lower $V_{\rm CG}$ to realize identical T_P. The difference between required $V_{\rm CG}$ under CHE and CHISEL operation increases as V_D is decreased. It is therefore evident from the above results that CHISEL programming remains efficient (low $V_{\rm CG}$ and/or V_D requirement) even for small $L_{\rm FG}$ cells. Note that higher $V_{\rm CG}$ increases power consumption (varies as square of $V_{\rm CG}$), while higher V_D accelerates drain disturbs and cycling induced degradation. These effects get aggravated as the cells are scaled. There-

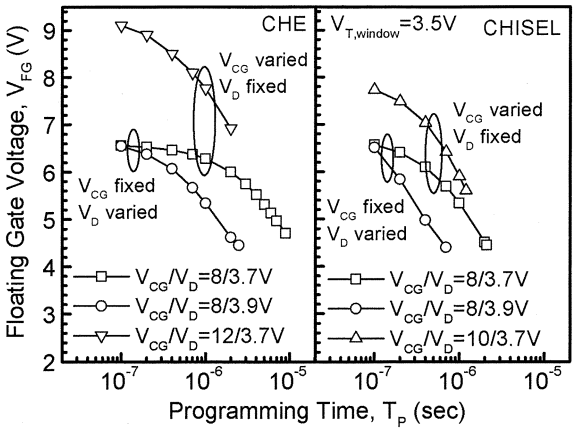


Fig. 5. FG voltage transients of a $L_{\rm FG} = 0.26\ \mu$m cell under CHE and CHISEL programming operation. $V_{\rm FG}$ was calculated from measured program V_T transients as a function of $V_{\rm CG}$ (V_D fixed at 3.7 V) and V_D ($V_{\rm CG}$ fixed at 8 V).

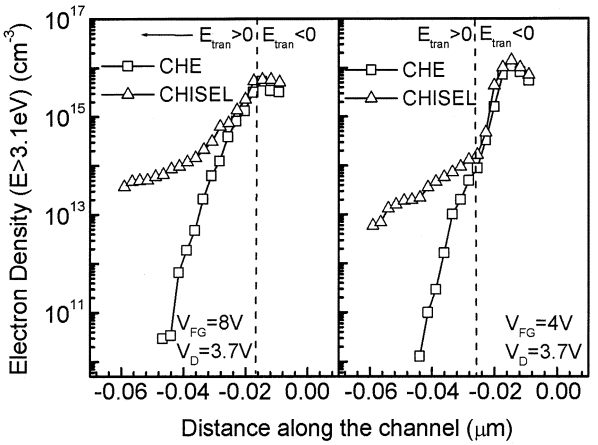


Fig. 6. Hot ($E > 3.1$ eV) electron density distribution along the channel, simulated near the interface of a $L_{\rm FG} = 0.26\ \mu$m FG-contacted device under CHE and CHISEL operation. The point in the channel where $E_{\rm TRAN}$ changes sign is shown by dotted line. Electron injection area is on the left of the dotted line ($E_{\rm TRAN} > 0$).

fore, CHISEL operation is a better option for low power programming of scaled NOR Flash EEPROMs.

We now explain the physical mechanism responsible for efficient CHISEL programming at low $V_{\rm CG}$ and V_T saturation under CfHE and CHISEL operation. As a first step it is necessary to estimate $V_{\rm FG}$ during programming, which is calculated using the following formula [10]:

$$\begin{aligned} V_{\rm FG}(t) =& \alpha_G.V_{\rm CG} + \alpha_D.V_D + \alpha_S.V_S \\ &+ \alpha_B.\varphi_{s,\rm av}(t) + \frac{Q_{\rm FG}(t)}{C_T} \\ Q_{\rm FG}(t) =& \Delta V_T(t).C_{\rm ONO} \end{aligned} \tag{1}$$

where $\Delta V_T(t)$ is the change in threshold voltage during programming from natural V_T of the cell, $\varphi_{s,\rm av}(t)$ is the average surface potential during programming (obtained from device simulation), $C_{\rm ONO}$ is the capacitance of IPD and C_T is the total capacitance. The values of coupling coefficients $(\alpha_G, \alpha_D, \alpha_S, \alpha_B)$ used for the calculation are 0.6, 0.05, 0.05 and 0.3 respectively. The value of α_G is calculated by sub-threshold slope technique [19]. α_D is calculated by measuring the change in $V_{\rm CG}$ with V_D for equivalent I_D and using the following formula:

$$\alpha_D = \alpha_G \left(\frac{|\Delta V_{\rm CG}|}{|\Delta V_D|} \right) \tag{2}$$

and after extracting α_D, α_B is calculated as follows:

$$\alpha_B = 1 - (\alpha_G + \alpha_D + \alpha_S) \text{ where } \alpha_S = \alpha_D. \tag{3}$$

Fig. 5 shows the calculated $V_{\rm FG}$ transients during CHE and CHISEL programming under different $V_{\rm CG}(V_D = 3.7\ \text{V})$ and $V_D(V_{\rm CG} = 8\ \text{V})$. Since $\Delta V_{\rm FG} = -\alpha_G \Delta V_T$, electron injection into FG during programming reduces $V_{\rm FG}$, and the reduction is faster for faster programming under higher V_D or higher $V_{\rm CG}$. However since $V_{\rm FG} = \alpha_G V_{\rm CG}$, $V_{\rm FG}$ drops down to a very low value for programming at higher V_D (low and moderate $V_{\rm CG}$), but always remains higher when programming is performed at higher $V_{\rm CG}$ (any V_D value).

Fig. 6 shows the effect of $V_{\rm FG}$ on hot electron density (having energy $E \leq 3.1$ eV) distribution, simulated at the interface and along the channel of a FG contacted device under CHE and CHISEL operation. The origin was chosen at FG edge near the drain, and the drain junction is located at about $-0.02\ \mu$m. It can be observed that under both high and low $V_{\rm CG}$, CHISEL operation broadens the hot electron distribution profile by increasing the hot electron density mainly in the channel region. The dotted line in Fig. 6 represents the $E_{\rm TRAN} = 0$ point in the channel. For a given $V_{\rm FG}$ and V_D, $E_{\rm TRAN} > 0$ region is toward the center of channel (left of the dotted line) and $E_{\rm TRAN} < 0$ region is toward the drain junction (right of the dotted line). Note that hot electrons injected from the portion of channel having $E_{\rm TRAN} > 0$ reach FG and contribute to V_T shift during programming. As programming continues, $V_{\rm FG}$ drops and the $E_{\rm TRAN} = 0$ point moves toward the center of the channel. Therefore, hot electron density in the $E_{\rm TRAN} > 0$ region reduces which slows down the programming transient (saturation). At high $V_{\rm FG}$, sufficient hot electron density exists in the $E_{\rm TRAN} > 0$ region for both CHE and CHISEL operation. Therefore both CHE and CHISEL operation remain efficient at high $V_{\rm CG}$. On the other hand at low $V_{\rm FG}$, hot electron density in the $E_{\rm TRAN} > 0$ region is much higher for CHISEL compared to CHE operation. Therefore unlike CHE operation, CHISEL programming at moderate $V_{\rm CG}$ does not slow down at longer times. At higher V_D, there is more electron injection during the initial period of programming. This reduces $V_{\rm FG}$ at a much faster rate and cause early V_T saturation. However, since hot electron density in the $E_{\rm TRAN} > 0$ region is much higher for CHISEL compared to CHE operation, V_T saturation takes place at higher V_T for the former compared to the latter.

B. Effect of L_FG Scaling

Fig. 7 shows the impact of $L_{\rm FG}$ scaling on CHE and CHISEL programming performance of type B cells. Fig. 7(a) shows T_P (for V_T shift of 3.5 V) as a function of $L_{\rm FG}$ at a constant V_D and $V_{\rm CG}$. Fig. 7(b) shows the V_D required to achieve $T_P = 3\ \mu$s (at $V_{\rm CG} = 8$ V) as a function of $L_{\rm FG}$. Compared to CHE, CHISEL programming always shows lower T_P and V_D for all $L_{\rm FG}$ values used in this study. For both CHE and CHISEL programming,

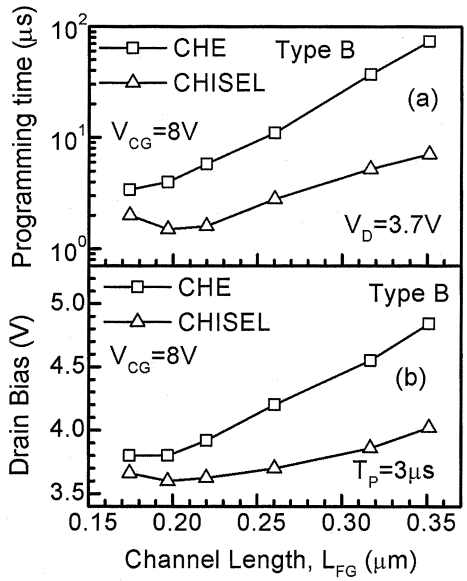


Fig. 7. Effect of FG length scaling on (a) programming time at $V_{\rm CG}/V_D = 8/3.7$ V and (b) drain bias required for $T_P = 3$ μs at $V_{\rm CG} = 8$ V under CHE and CHISEL operation. T_P was calculated for a 3.5 V program V_T shift.

T_P (at fixed V_D, $V_{\rm CG}$) and V_D (at fixed T_P, $V_{\rm CG}$) reduce with decrease in $L_{\rm FG}$. However the rate of reduction is lower for CHISEL programming, which shows a turn around at smaller $L_{\rm FG}$ values. The reduction of CHISEL programming efficiency at small $L_{\rm FG}$ cells is consistent with previous results [5], [16], [17].

C. Effect of Technology Parameters

Note that CHISEL programming efficiency is determined by secondary impact ionization, which in-turn is governed by $E_{\rm TRAN}$ near the drain junction [1]–[9]. As $L_{\rm FG}$ is scaled (keeping all other cell parameters constant), short-channel effect (SCE) comes into play that reduces $E_{\rm TRAN}$, secondary impact ionization and CHISEL programming efficiency [5], [16], [17]. Therefore to restore CHISEL programming efficiency at short $L_{\rm FG}$ cells it is necessary to increase $E_{\rm TRAN}$ near the drain junction. As mentioned in Section II, four different (A, B, C, and D) types of cells were fabricated to study the impact of technological parameters on CHISEL programming efficiency. Compared to type A cells, type B cells have higher channel implant, type C cells have higher channel implant and shallower S/D junction while type D cells have halo doping around the S/D junction. Therefore, compared to type A cells, the doping around the S/D junction of all the other (type B, C and D) cells are made more abrupt, which will give rise to higher $E_{\rm TRAN}$ under identical V_D and V_B during programming operation.

Fig. 8 shows T_P as a function of $V_{\rm CG}$ and V_D measured under CHISEL programming of $L_{\rm FG} = 0.26$ μm type A through D cells. As shown, for any combination of $V_{\rm CG}$ and V_D type D cells (halo) show fastest T_P while type C cells show faster T_P compared to type A and B cells (type A through C do not have halo). Note that the improvement in T_P (type D versus type C versus type B versus type A) is more at higher $V_{\rm CG}$ but insensitive to changes in V_D.

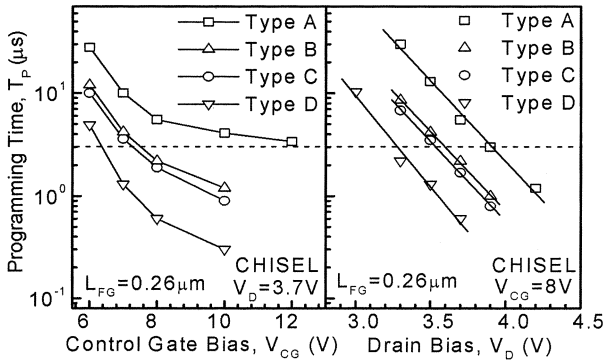


Fig. 8. Effect of control gate and drain bias on programming time for $L_{\rm FG} = 0.26$ μm type A, B, C, and D cells under CHISEL operation. T_P was calculated for a 3.5 V program V_T shift as a function of $V_{\rm CG}$ (V_D fixed at 3.7 V) and V_D ($V_{\rm CG}$ fixed at 8 V).

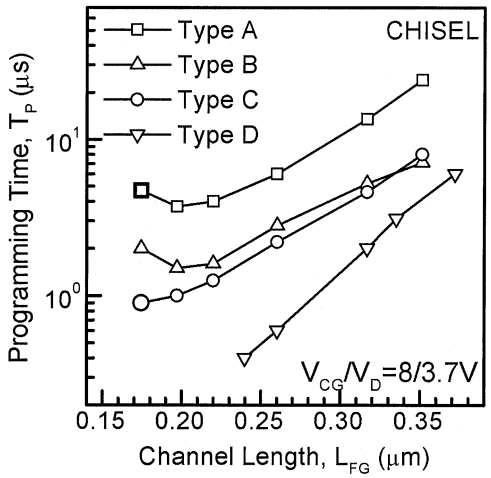


Fig. 9. Programming time as a function of FG length for type A, B, C, D Flash cells under CHISEL operation at $V_{\rm CG}/V_D = 8/3.7$ V. T_P was calculated for a 3.5 V program V_T shift.

Fig. 9 shows T_P of type A through D cells as a function of $L_{\rm FG}$, obtained under CHISEL operation at a fixed $V_{\rm CG}$ and V_D. As shown, the programming efficiency is improved and no T_P turn around is seen at smaller $L_{\rm FG}$ cells optimized with higher channel doping and lower S/D junction depth (type C). However, maximum improvement is achieved for halo implantation (type D), which shows the steepest T_P versus $L_{\rm FG}$ slope when compared to other (type A through C) optimized cells.

D. Impact of Technological Parameters on Drain Disturb

Drain disturb is an important concern for CHISEL operation. This is due to higher voltage drop across the drain junction ($V_D + |V_B|$) that increases band-to-band tunneling (BBT) and a generation of hot electrons and hot holes [5], [18]. Injection of hot electrons or hot holes into the FG cause charge gain (in erased cell) or charge loss (in programmed cell) respectively. All the parameter optimization (higher channel doping, halo implant, smaller S/D junction depth) required to improve CHISEL programming efficiency of small $L_{\rm FG}$ cells also results in higher $E_{\rm TRAN}$ and therefore higher drain disturb.

Table II shows the junction breakdown voltage ($V_{\rm BD}$) and leakage current (in programmed state at maximum V_D) for

TABLE II
THE JUNCTION BREAKDOWN VOLTAGE AND LEAKAGE CURRENT (IN PROGRAMMED STATE WITH $V_T = 5.4$ V AND MAXIMUM V_D) FOR 0.26 μm TYPE A, B, C AND D FLASH CELLS. MEASUREMENTS WERE PERFORMED ON EQUIVALENT FLOATING GATE-CONTACTED DEVICES. I_{leak} IS MEASURED FOR V_B OF -2 V

Type of the cell	L_{FG} (μm)	V_{BD} (V)	I_{leak} (pA)
A	0.26	7.6	125.4
B	0.26	7.3	166.1
C	0.26	7.0	588.7
D	0.26	6.6	1620

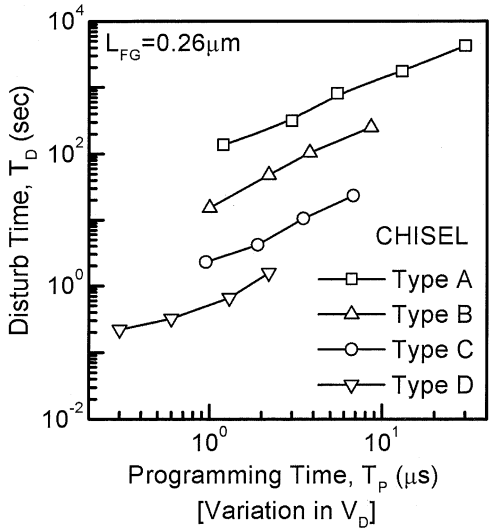


Fig. 10. Disturb time as a function of programming time for $L_{\text{FG}} = 0.26\ \mu$m type A, B, C, D Flash cells under CHISEL operation. T_P was calculated for a 3.5 V program V_T shift. T_D was calculated for a 0.1 V V_T shift.

$L_{\text{FG}} = 0.26\ \mu$m type A, B, C, and D cells. The consistent decrease in V_{BD} from type A to D cells is expected due to the increase in the abruptness of the S/D junction. However for all the cells, V_{BD} is well above maximum $V_D + |V_B|$ used in this study. Due to this margin in applied voltage to V_{BD}, the junction leakage current is also within tolerable limits.

Fig. 10 shows the disturb time T_D as a function of T_P (achieved by varying V_D) for $L_{\text{FG}} = 0.26\ \mu$m type A, B, C, and D cells. T_D is defined as the time required for a V_T change of 0.1 V during charge gain drain disturb measurements [5], [18]. When compared to type A cells under constant T_P, T_D decreases by 1.5 orders of magnitude for type C cells and by over two orders of magnitude for type D cells. It is important to note that the changes in technological parameters, which causes maximum increase of programming efficiency (fastest T_P), also causes maximum increase in drain disturb (lowest T_D). This is expected since both T_P and T_D under CHISEL programming is proportional to E_{TRAN}. However, the degradation in T_D/T_P is worst for type D cells compared to type B and C cells. This is because halo implantation makes the S/D junction very abrupt, which increases BBT and hence reduces T_D much more than other cells. Therefore, CHISEL

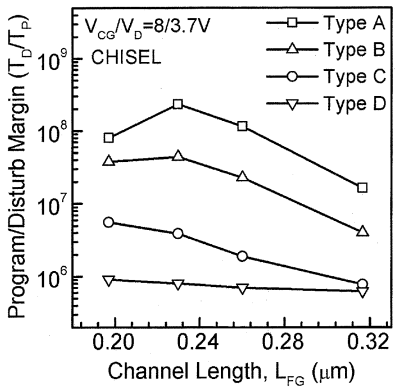


Fig. 11. Program/disturb margin as a function of FG length for type A, B, C, D Flash cells under CHISEL operation at $V_{\text{CG}}/V_D = 8/3.7$ V. T_P was calculated for a 3.5 V program V_T shift. T_D was calculated for a 0.1 V V_T shift.

efficiency optimization has to be performed by keeping the program/disturb margin (T_D/T_P) under acceptable limits.

Fig. 11 shows T_D/T_P as a function of L_{FG} for type A, B, C, and D cells. When compared to type A cells for a given L_{FG}, T_D/T_P degrades for type B and C cells and even more for type D cells. For no-halo (type A, B and C) cells, T_D is largely insensitive while T_P decreases with decrease in L_{FG}, therefore T_D/T_P increases with decreases in L_{FG}. Note that the T_D/T_P turn around at lower L_{FG} for type A and B cells is due to the turn around in T_P as shown in Fig. 7. On the other hand for halo (type D) cells there is not much improvement in T_D/T_P as L_{FG} is decreased. Therefore, though halo implantation is a better option for improving CHISEL programming efficiency for smaller L_{FG} cells, it comes with a penalty of much degraded program/disturb margin. Note that the worst-case program/disturb margin should be in the range of 10^6–10^7 (depending on number of cells/bit line, cycling requirements, etc.), larger than that obtainable with halo cells. On the other hand, CHISEL programming efficiency improvement by using higher channel doping and lower S/D junction depth seems to be a better option since it also keeps program/disturb margin within reasonable limits especially at smaller L_{FG} cells.

IV. CONCLUSION

To summarize, we have studied the effect of control gate (V_{CG}) and drain (V_D) biases, FG length (L_{FG}) and technological parameters (channel doping, halo implant, S/D junction depth) on CHISEL programming efficiency of NOR Flash EEPROM cells.

CHISEL operation offers faster programming time (T_P) at an identical bias (V_{CG}, V_D) and requires lower V_{CG} and/or V_D for similar T_P when compared against conventional CHE operation. This higher programming efficiency of CHISEL operation has been verified on small L_{FG} cells and over a wide range of bias choices. Using full band Monte-Carlo simulations we show that CHISEL operation results in a broader hot electron distribution profile, which can explain the higher efficiency of CHISEL programming at lower biases. The lower V_{CG} and V_D requirement for CHISEL operation makes it an ideal low-power, reliable programming scheme for NOR Flash EEPROMs.

Consistent with previous reports, CHISEL programming efficiency is shown to reduce for smaller $L_{\rm FG}$ cells. Suitable technology parameters are modified (increased channel doping, decreased S/D junction depth or added halo implant) to improve the programming efficiency of smaller $L_{\rm FG}$ cells. However, all these parameter changes are shown to increase drain disturb and degrade program/disturb margin. It is shown that the improvement in CHISEL programming efficiency at smaller $L_{\rm FG}$ is highest for halo-implanted cells but it also suffers from severe reduction in program/disturb margin. However, nonhalo cells having higher channel doping and lower S/D junction depth offers good improvement in CHISEL programming efficiency without severely degrading the program/disturb margin especially for smaller $L_{\rm FG}$ cells. We conclude that by judicious choice of technological parameters high CHISEL programming efficiency can be maintained for smaller $L_{\rm FG}$ cells.

REFERENCES

[1] J. D. Bude, A. Frommer, M. R. Pinto, and G. R. Weber, "EEPROM/Flash sub-3.0 V drain-source bias hot carrier writing," in *IEDM Tech. Dig.*, 1995, pp. 989–992.

[2] J. D. Bude *et al.*, "Secondary electron Flash—A high performance low power Flash technology for 0.35 μm and below," in *IEDM Tech. Dig.*, 1997, pp. 279–282.

[3] M. Mastrapasqua, "Low voltage Flash memory by use of a substrate bias," *Microelectron. Engineering*, vol. 48, pp. 389–394, 1999.

[4] D. Esseni, A. D. Strada, P. Cappelletti, and B. Ricco, "A new and flexible scheme for hot-electron programming of nonvolatile memory cells," *IEEE Trans. Electron Devices*, vol. 46, pp. 125–133, Jan. 1999.

[5] S. Mahapatra, S. Shukuri, and J. D. Bude, "CHISEL Flash EEPROM—Part-1: Performance and scaling," *IEEE Trans. Electron Devices*, vol. 49, pp. 1296–1301, July 2002.

[6] J. D. Bude, "Gate current by impact ionization feedback in sub-micron MOSFET technologies," in *Proc. Symp. VLSI Technol.*, 1995, pp. 101–102.

[7] D. Esseni and L. Selmi, "A better understanding of substrate enhanced gate current in MOSFETs and Flash cells—Part I: Phenomenological aspects," *IEEE Trans. Electron Devices*, vol. 46, pp. 369–375, Feb. 1999.

[8] L. Selmi and D. Esseni, "A better understanding of substrate enhanced gate current in MOSFETs and Flash cells—Part II: Physical analysis," *IEEE Trans. Electron Devices*, vol. 46, pp. 376–382, Feb. 1999.

[9] J. D. Bude, M. R. Pinto, and R. K. Smith, "Monte carlo simulation of CHISEL Flash memory cell," *IEEE Trans. Electron Devices*, vol. 47, pp. 1873–1881, Oct. 2000.

[10] P. Pavan and R. Bez, "The industry standard Flash memory cell," in *Flash Memories*, P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, Eds. Boston, MA: Kluwer, 1999.

[11] C. Y. Hu *et al.*, "A convergence scheme for over erased Flash EEPROMs using substrate enhanced hot electron injection," *IEEE Electron Device Lett.*, vol. 11, pp. 500–502, Nov. 1995.

[12] ——, "Substrate-current-induced hot electron (SCIHE) injection: A new convergence scheme for Flash memory," in *IEDM Tech. Dig.*, 1995, pp. 283–286.

[13] K. Yoshikawa *et al.*, "Comparison of current Flash EEPROM erasing methods: Stability and how to control," in *IEDM Tech. Dig.*, 1992, pp. 595–598.

[14] S. Mahapatra, S. Shukuri, and J. D. Bude, "CHISEL Flash EEPROM—Part-2: Reliability," *IEEE Trans. Electron Devices*, vol. 49, pp. 1302–1307, July 2002.

[15] N. R. Mohapatra, S. Mahapatra, V. R. Rao, S. Shukuri, and J. D. Bude, "Effect of programming biases on the reliability of CHE and CHISEL Flash EEPROMs," in *IRPS Tech. Dig.*, 2003, pp. 518–522.

[16] D. Esseni, L. Selmi, A. Ghetti, and E. Sangiorgi, "Injection efficiency of CHISEL gate currents in short MOS devices: Physical mechanisms, device implications and sensitivity to technological parameters," *IEEE Trans. Electron Devices*, vol. 47, pp. 2194–2200, Nov. 1999.

[17] ——, "The scaling properties of CHISEL and CHE injection efficiency in MOSFETs and Flash memory cells," in *IEDM Tech. Dig.*, 1999, pp. 275–278.

[18] D. R. Nair, N. R. Mohapatra, S. Mahapatra, S. Shukuri, and J. Bude, "The effect of CHE and CHISEL programming operation on drain disturb in Flash EEPROMs," in 10th International Symposium on the Physical and Failure Analysis of Integrated Circuits, 2003, pp. 164–167.

[19] M. Wong, D. K.-Y. Liu, and S. S.-W. Huang, "Analysis of the subthreshold slope and linear transconductance techniques for the extraction of the capacitance coupling coefficients of floating-gate devices," *IEEE Electron Device Lett.*, pp. 566–568, Nov. 1992.

Nihar R. Mohapatra (S '01) received the B. Tech. degree in electrical engineering from Sambalpur University, Orissa, India in 1998. Since July 1999, he has been pursuing the Ph.D. degree at the Indian Institute of Technology (IIT), Bombay, India.

His current interests are in the areas of MOS physics and technology, characterization, modeling and simulation. He has worked on modeling and simulation of high-K gate dielectrics, short-channel effects and hot-carrier reliability in MOS transistors and Flash memories.

Deleep R. Nair (S '98) was born in New Delhi, India, in 1978. He received the B.Tech. degree in electronics and communication engineering from REC Calicut, India in 1999, and the M.Tech. degree in electrical engineering from the Indian Institute of Technology (IIT), Bombay, India, in 2001. Since 2001, he has been pursuing the Ph.D. degree at IIT, Bombay.

His current interests include MOS physics and technology, characterization and numerical simulation of semiconductor devices. He has worked on the characterization and simulation of Flash memories and numerical modeling of quantum effects in MOS devices.

S. Mahapatra received the M.Sc. degree in physics from Jadavpur University, Calcutta, India and the Ph.D. degree in electrical engineering (microelectronics) from the Indian Institute of Technology (IIT), Bombay, India, in 1995 and 1999 respectively. His doctoral thesis was on the study of hot-carrier degradation in conventional, channel engineered and high-k MOSFETs using a novel charge pumping technique.

From 2000 to 2001 he was at Bell Laboratories, Lucent Technologies, Murray Hill, NJ. At Bell Labs, he played a key role in designing and developing the unit cell of the world's first commercial CHISEL Flash memory, and was also involved in studies of interface characterization of GaAs-GdO FETs, hot-carrier instability in RFLDMOS and p-MOSFET bias temperature instability. Since January 2002 he has been with the IIT Department of Electrical Engineering, where he is presently an Assistant Professor. His present research interest involves semiconductor device physics, simulation, modeling and characterization, novel devices, hot-carrier and bias temperature reliability issues in MOSFETs, Flash memories and high-k gate dielectrics. He has published more than 30 papers in refereed international journals and conferences, and worked as a reviewer for many international journals and conferences.

V. Ramgopal Rao (M'98–SM'02) received the M.Tech. from the Indian Institute of Technology (IIT), Bombay, India, in 1991 and Dr.-Ing. (magna cum laude) degree from the Faculty of Electrical Engineering, Universitäet der Bundeswehr, Munich, Germany, in 1997. His doctoral thesis was on planar-doped-barrier sub-100-nm channel-length MOSFETs.

He was a Deutscher Akademischer Austauschdienst (DAAD) Fellow for three years during 1994 to 1996, and again from February 1997 to July 1998, and in 2001, a Visiting Scholar with the Electrical Engineering Department, University of California, Los Angeles, CA. He is currently an Associate Professor in the Department of Electrical Engineering, IIT Bombay. His areas of interest include physics, technology and characterization of short-channel MOSFETs, CMOS scaling for mixed signal applications, and bio-MEMS. He has over 120 publications in these areas in refereed international journals and conference proceedings and holds couple of patents in these areas. He is an organizing committee member for the various international conferences held in India.

Dr. Rao is a Fellow of IETE and the chairman of the IEEE AP/ED Bombay Chapter.

S. Shukuri was born in Oita, Japan, in 1958. He received the B.S. degree in electrical engineering from Yamanashi University, Yamanashi, Japan, in 1980 and the M.S. degree in electrical engineering from Kyushu University, Fukuoka, Japan, in 1982.

In 1982, he joined Central Research Laboratory, Hitachi Ltd., Tokyo, Japan, where he engaged in the research of the focused ion beam implantation and its application to LSIs. From 1987 to 1992, he has engaged in the process and device design of high-speed BiCMOS and DRAM cell. In 1993 he joined Semiconductor & Integrated Circuits Div., Hitachi Ltd., where he has been in charge of Flash memory device development.

Mr. Shukuri is a member of the Japan Society of Applied Physics.

Jeff D. Bude was born in St. Louis, MO on July 26, 1966. He received the B.S., M.S., and the Ph.D. degrees in electrical engineering (with honors) at the University of Illinois, Urbana-Champaign, in 1987, 1989, and 1992, respectively.

In 1992 he joined Bell Laboratories-Lucent Technologies, Murray Hill, NJ, as Member of the Technical Staff. and became Distinguished Member of the Technical Staff in 1999. In 2000, he became director of the High Speed Electronics Device Research Department, Agere Systems, Allentown, PA. His research interests are mainly focused on transistor and nonvolatile memory device physics. He has been involved in the simulation and design of high-speed devices and in research emphasizing hot carrier effects and reliability.