Extended Abstracts of the 1994 International Conference on Solid State Devices and Materials, Yokohama, 1994, pp. 518-520

New Scaling Scenario for Channel Hot Electron Type Flash EEPROM

S. Ueno, H. Oda, N. Ajika, M. Inuishi and N. Tsubouchi

ULSI Laboratory, Mitsubishi Electric Corporation

4-1 Mizuhara, Itami, Hyogo 664 JAPAN

The new scaling rule of flash EEPROM is presented to keep the programming current constant. In the new scaling rule, the concept of PBC (Programming Best Condition) is introduced to discuss the programming time by using the maximum gate current. Using PBC concept, it is clarified that there exists the scaling limitation for the drain, the gate and the swing voltage. Moreover it is derived that the supply voltage after scaling should be reduced by the scaled difference between the voltage before scaling and the limitation, to keep the programming current constant.

[Introduction]

Many scaling theories of MOSFETs have been evolved to improve performance and packing density. However scaling methodology of flash EEPROM has not been thoroughly discussed, yet[1]. In this paper, the new scaling scenario is proposed, including the concept of PBC (Programming Best Condition). PBC means the applied voltage condition where the programming time becomes shortest. In this scenario, there exists the scaling limitation for the drain, the gate and the swing voltage. So the voltages can not be simply scaled by $1/k$ according to the scaling factor k (>1) . Instead, the voltage shift from the limitation should be scaled to keep the programming current constant.

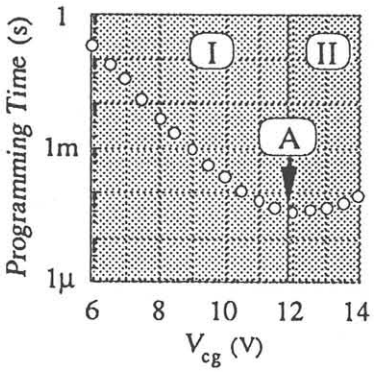

Fig. 1: *Programming time* dependence on V_{cg}.

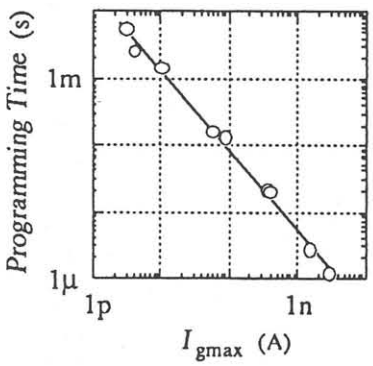

Fig. 2: The relationhsip between *Programming time* and I_{gmax} under PBC.

[Programming Best Condition (PBC)]

The programming time is analytically calculated from equation 1 which expresses charging the capacitance.

$$\textit{Programming Time} = C_t \int_{V_{end}}^{V_{start}} \frac{dV_{fg}}{I_g(V_{fg})} \qquad (1)$$

The charge balance equation 2 in the EPROM structure is given by

$$Q_{fg} = C_{fc}(V_{fg} - V_{cg}) + C_{fd}(V_{fg} - V_d) + C_{fs}(V_{fg} - V_s) + C_{fb}(V_{fg} - V_s - V_{th}) \qquad (2)$$

Figure 1 shows the programming time versus the control gate voltage using these equations 1 and 2. In the region I of figure 1, the programming time reduces with increasing control gate voltage, since the gate current increases due to the oxide field favoring the hot carrier injection. On the other hand, in the region II, the programming time increases, because the gate current reduces due to decrease in the electric field near the drain junction. There exists minimum programming time for a certain gate bias as shown figure 1. This is due to the largest gate current for this bias condition. We call this PBC

(Programming Best Condition) . It should be noted that the programming time under PBC is exponentially proportional to the maximum gate current as shown in figure 2. Hence, the scaling scenario under PBC can be discussed in terms of the maximum gate current with respect to the drain voltage and the impurity concentration. Moreover, under PBC (I_{gmax} condition) , the gate voltage is linearly related to the drain voltage. Therefore, the gate voltage is scaled, when the drain voltage is reduced.

[Gate Current Characteristics]

At first, we have to clarify the relationship between the maximum gate current and the drain voltage or the impurity concentrations for discussing the scaling scenario. Based on the lucky-electron model [2], the gate current can be expressed as

$$I_g = I_d \, P(E_{ox}) \left(\frac{\phi_b E_m}{\lambda}\right)^2 exp\left(\frac{\lambda}{\phi_b E_m}\right) \qquad (3)$$

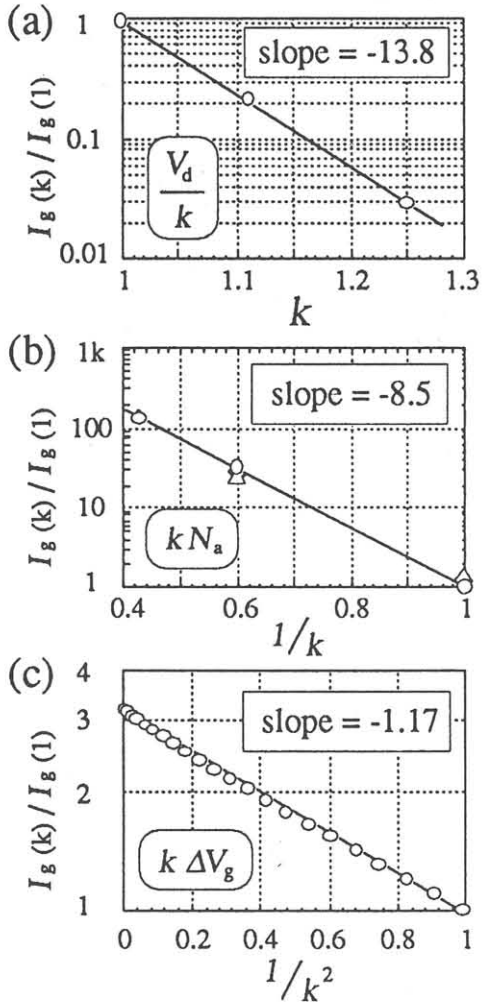

Fig. 3: The sensitivity of I_{gmax} to the (a) drain voltage V_d, (b) impurity concentrations N_a and (c) the shift of gate voltage ΔV_g.

where λ is the optical phonon mean free pass and ϕ_b is the effective barrier height. E_m is the maximum lateral electric field, which is derived using quasi-two dimensional analysis as follows.

$$E_{max} = \frac{V_d - V_{dsat}(N_a)}{\ell} \qquad (4)$$

Where V_{dsat} is the saturation drain voltage, which is proportional to N_a by solving the Poisson's equation at velocity saturation region [3]. Figure 3(a) shows the normalized I_{gmax} ($I_g(k)/I_g(1)$), when the drain voltage is scaled by k (V_d/k). The I_{gmax} is proportional to k, because the gate current is proportional to $1/V_d$. And it is derived that the sensitivity of the I_{gmax} (SV_d), when the V_d is scaled, is expressed as follows by the least square fitting in figure 3(a).

$$log\left(\frac{I_{gmax(k)}}{I_{gmax(1)}}\right) = S_{V_d} = -13.8\,(k-1) \qquad (5)$$

Figure 3(b) shows that the I_{gmax} is proportional to $1/k$, when the impurity concentration is scaled as kN_a. This means that the gate current is proportional to $1/N_a$ and the sensitivity of the I_{gmax} against the scaled N_a (SN_a) is predicted by equation 6.

$$log\left(\frac{I_{gmax(k)}}{I_{gmax(1)}}\right) = S_{N_a} = -8.5\left(\frac{1}{k} - 1\right) \qquad (6)$$

Moreover, it is necessary to include the sensitivity of the I_g to the gate voltage shift (ΔV_g) from the I_{gmax} condition, to discuss the scaling of the gate swing between the high and the low states. Figure 3(c) shows the normalized I_g, when the swing of the gate voltage scaled by k ($\Delta V_g/k$). The sensitivity of the I_g to the scaled $\Delta V_g/k$ ($S\Delta V_g$) is predicted from figure 3(c) as follows.

$$log\left(\frac{I_{g\ at\Delta V_g(k)}}{I_{g\ at\Delta V_g(1)}}\right) = S_{\Delta V_g} = -1.17\left(\frac{1}{k^2} - 1\right) \qquad (7)$$

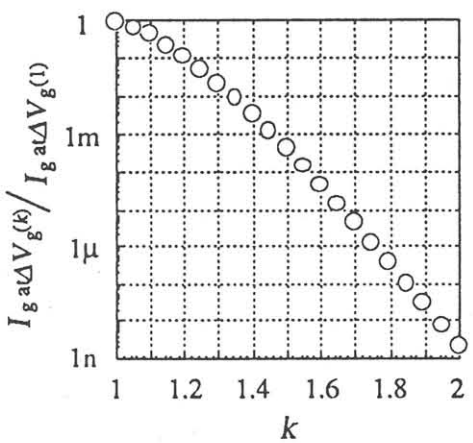

Fig. 4: I_{gmax} versus scaling of V_d, N_a and ΔV_g.

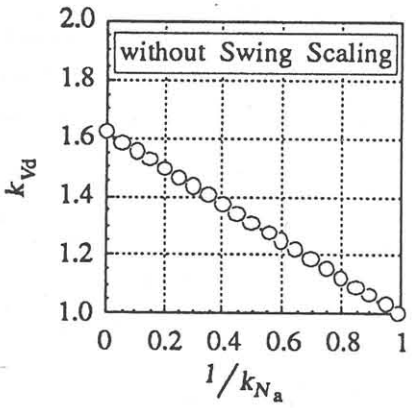

Fig 5: The relationship between k_{Na} and k_{Vd} for keeping the programming current constant, when the gate swing is not scaled down.

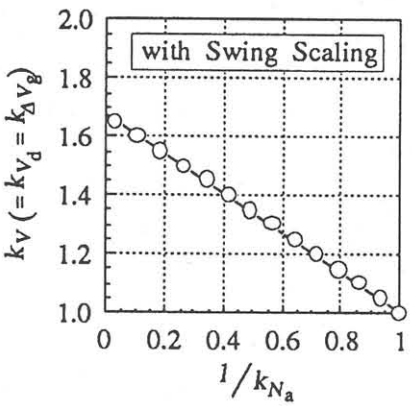

Fig 6: The relationship between k_{Na} and k_V ($=k_{Vd}=k_{V\Delta Vg}$) for keeping the programming current constant, when the gate swing is scaled down.

Table I : Summarize the new scaling law

Gate Swing	*keeping*	*scaleing*
Impulity Concentration	k	k
Supply Voltage at Writing Mode		
$V_d - V_{d.limit}$	$1/k$	$1/k$
$V_{fg} - V_{fg.limit}$	$1/k$	$1/k$
$\Delta V_g - \Delta V_{g.limit}$	/	$1/k$
Gate Current	1	1
Programming Time	$1/k$	$1/k$
$V_{d.limit}$	3.1	3.0
$V_{fg.limit}$	$1.5 + 1/2\ V_{fg.swing} + V_{th}$	
$\Delta V_{g.limit}$	/	1.2

[Scaling Scenario]

In the view of the above considerations, the change of the maximum gate current can be predicted by the multiplied product of V_d, N_a and ΔV_g sensitivities. Figure 4 shows the relation between scaling factor and the maximum gate current. Note that no scaling can keep the gate current and the programming ability constant. This is because, the sensitivity for the drain voltage is much higher than that for the others. Therefore, to keep gate current constant, the scaling for V_d should be optimized in combination with the other scaling factors.

From this point of view, the new scaling law to keep the gate current constant, is deduced by keeping the summation of sensitivity factors equal to zero as can be seen from equation 5 to 7. For example, when the drain voltage and the impurity concentrations are scaled except the swing voltage, the relation of equation 8 should be held.

$$S_{V_d} + S_{N_a} = -13.8\,(k_{V_d} - 1) - 8.5\,(\frac{1}{k_{N_a}} - 1) = 0 \qquad (8)$$

where k_{Vd} and k_{Na} are the scaling factors of the drain voltage (under PBC) and that of the impurity concentration. Figure 5 shows the relationship between the scaling factor for the impurity concentration (k_{Na}) and that for the drain voltage (k_{Vd}) from equation 8. k_{Vd} is proportionally to $1/k_{Na}$ and the least square fitted curve to the data intersect to the k_{Vd} axis at 1.6 ($1/k_{Na} = 0$, N_a = infinity). This indicates that the drain voltage can not be reduced below 3.1V($=V_{dlimit}$), even if the impurity concentration were infinity. Assuming the k_{Vd} is approximated as the linear function of k_{Na}, the drain voltage should be scaled as expressed by equation 9

$$V_d(k) = V_d(1) - \frac{V_d(1) - V_{d.limit}}{k} \qquad (9)$$

where k is the scaling factor, and the impurity concentration is $k \cdot N_a$. Moreover when the gate swing is scaled, the relationship between the scaling factor of the supply voltage (kv) and that of the impurity concentration is given by equation 10 and figure 6.

$$S_{V_d} + S_{N_a} + S_{\Delta V_g} = 0 \qquad (10)$$

In the same way as previously discussed, the limitation of the drain voltage and that of the gate swing are obtained to be 3.0V and 1.2V respectively. Hence the supply voltage can be reduced by (V - V_{limit}) / k, using the above limitation voltages.

The new scaling scenario is summarized in Table I. Comparing both of the scaling scenarios, it can be seen that scaling of the gate swing is less effective in reducing the limitation voltage. This is because the gate current hardly increases, when the V_g is varied near the maximum gate current region.

[Conclusion]

The concept of Programming Best Condition (PBC) is used to discuss the effect of the scaling of V_d, V_g and V_g swing on the programming time, in proportional to I_{gmax}. Under PBC, there exists the limitation voltage for scaling. Each supply voltage can be reduced by the scaled difference from each limitation voltage, to keep the programming current constant.

1) K.Yoshikawa, et al, VLSI Tech. Symposium, (1991), p.79

2) S.Tam, et al, IEEE Trans. Electron Devices, ED-31, (1984), p.1116

3) N.G.Einspruch, Advanced MOS Device Physics, Academic Press, San Diego, (1989), p.130