Extended Abstracts of the 1996 International Conference on Solid State Devices and Materials, Yokohama, 1996, pp. 776-778

Optimum Voltage Scaling and Structure Design for the Low Voltage Operation of FN Type Flash EEPROM with High Reliability and Constant Programming Time

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

ULSI Laboratory, Mitsubishi Electric Corporation
4-1 Mizuhara, Itami, Hyogo 664 JAPAN Tel. 0727-84-7322, Fax. 0727-80-2693

A voltage scaling methodology has been developed by simple equations to keep the program time constant with improving the cell disturbance and the tunnel oxide reliability, while maintaining the constant tunnel oxide thickness. Using this voltage scaling, the optimum structure for low voltage operation can be investigated. As a result, it is found that the high coupling flash memories programmed by the whole channel FN tunneling are suitable for the low voltage operation and for high reliable flash memories.

[Introduction]

The reliability problems of flash memories are mainly due to high electric field of the tunnel oxide. This is because the high stress brings a serious degradation of silicon di-oxide films, especially in thin films. Therefore the decrease in the maximum electric field of the SiO2 with the constant tunnel oxide thickness is the key issue to achieve the highly reliable flash memories. Moreover low voltage operation of flash EEPROMs becomes important increasingly with increase in the demand of portable instruments. High coupling ratio flash devices have been reported (1,2) for lowering the supply voltage. However the electric field for the disturbed cells becomes high with increase in the coupling ratio by the reported methods (1,2). In this paper, we present the voltage scaling rules for keeping the programming time constant with a decrease in the maximum electric field while maintaining the constant tunnel oxide thickness, to realize the highly reliable flash memories. This rule can be developed by expressing the programming time dependence of the gate/drain structure by simple analytical equations. Moreover it is clarified that the high coupling cells with the whole channel programming operation are suitable for the high reliability and the low voltage operation of the FN type flash memories.

[Samples]

The flash memory cells studied in this work use the Fowler Nordheim (FN) tunneling current for both the electron injection (erasing) and the extraction (programming) where we call the erased Vth "high level" and the programmed Vth "low level". The source and the control gate voltages are divided to optimize the cell disturbance immunity.

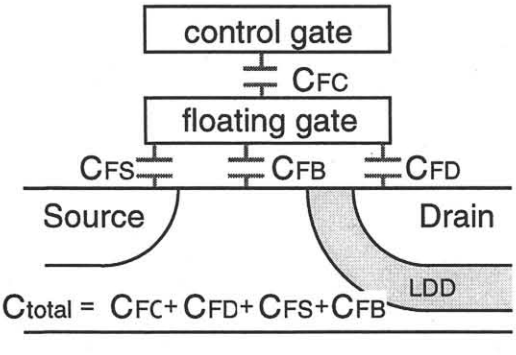

Fig. 1: samples

[Voltage Scaling Methodology]

In our previous work [3], the programming time is expressed by integrating the 1/IFG from the high level floating gate voltage (VFG,high) to the low level floating gate voltage (VFG,low) as in equation 1.

$$Prog.\ Time = \int_{V_{FG,high}}^{V_{FG,low}} \frac{C_{total}}{I_{FG}(V_{FG})}\, dV_{FG} \qquad (1)$$

where the Ctotal is the total capacitance as in figure 1 and the IFG is the gate current characteristic as a function of the floating gate voltage. Assuming the FN current operation, equation 1 can be solved as shown in equation 2.

$$Prog.\ Time = \frac{C_{total}\ t_{OX}}{S_{FN}\ A_{FN}\ B_{FN}}\ exp\left(\frac{B_{FN}}{E_{OX,low}}\right) \times \left[\ 1 - exp\left(\frac{B_{FN}}{E_{OX,high}} - \frac{B_{FN}}{E_{OX,low}}\right)\ \right] \qquad (2)$$

where the AFN and the BFN are the fitting parameters derived from the FN-plot. SFN is the tunneling area of the FN current. The Eox,high and the Eox,low are the tunnel oxide electric field at the high and the low levels. It is found that the programming time is mainly controlled by the low level electric field of the tunnel oxide. Therefore the voltage scaling methodology is derived by satisfying

$$\frac{C_{total}}{C_{FD}}\ exp\left(\frac{B_{FN}}{E_{OX,low}}\right) = constant \qquad (3)$$

Moreover, the drain and the control gate voltages are expressed as shown in equations 4 and 5, by satisfying both of the conditions for keeping the electric field constant and for

optimizing the cell disturbance.

$$\frac{C_{FC}}{C_{total}} V_{CG} = \frac{1}{2}\{ E_{OX,max}\, t_{OX} + \frac{Q_{FG}}{C_{total}} + a \} \quad (4)$$

$$(\frac{C_{FD}}{C_{total}} - 1)\, V_D = \frac{1}{2}\{ E_{OX,max}\, t_{OX} + \frac{Q_{FG}}{C_{total}} + a \} \quad (5)$$

where the Q_{FG} is the electric charge on the floating gate and α is the function of the voltage supplied at unselected cells.

These equations 4 and 5 indicate how to set the supply voltage for programming or reading and the threshold voltage at the high or the low level. Table 1 shows the voltage scaling methodology to maintain the constant programming time with reducing the tunnel oxide electric field. In this rule, the k_{FC}, k_{FD} and k_{FD}' are the structure dependent scaling factors. Moreover the k_E is the scaling factor for the tunnel oxide electric field, expressed as shown in equation 6 by solving the equation 3.

$$k_E = \frac{[E_{OX,low}]}{B_{FN}} \ln(k_{FD}) + 1 \quad (6)$$

The supply voltages of the V_{CG} and the V_D for programming must be scaled by $1/(k_{FC} \times k_E)$ and $1/k_{FD}$. The high and the low levels must be scaled by $1/(k_{FC} \times k_E)$ and $1/k_{FC}$ respectively.

Table 1: Voltage scaling rules for constant Etunnel,ox and for the constant program time.

Dimension		
Tunnel Oxide Thickness		1
Capacitance Coupling	FG - CG	k_{FC}
	FG - D	k_{FD}
$1-C_{FD}/C_{total}$		k_{FD}'
Voltage		
Sense Level	high	$1/(k_{FC} \cdot k_E)$
	low	$1/(k_{FC})$
Prog. Voltage	at Control Gate	$1/(k_{FC} \cdot k_E)$
	at Drain	$1/(k_{FD}')$
Read Voltage	at Control Gate	$1/(k_{FC} \cdot k_E)$
	at Drain	$1/(k_{FD} \cdot k_E)$
Scaling Result		
Q_{FG} dependence on V_{FG}	at high	$\sim 1/k_E$
	at low	~ 1
tunnel oxide Electric Field	at high	$1/k_E$
	at low	1
Program Time		1
Cell Distarbance immunity		-------

[Optimum structure for low voltage operation]

Suitable structure of the flash EEPROMs is demonstrated for low voltage operation. An increase in the C_{FC}/C_{total} tends to decrease the supply voltage. Figure 2 shows the relation between the V_D and the V_{CG} with respect to the drain overlap ratio and the coupling ratio for keeping the programming time = 1ms, based on the previous voltage scaling methodology. It is found that an increase in the C_{FC}/C_{total} makes the remarkable decrease in the supply voltage, especially in the V_{CG}. However an increase in the C_{FC}/C_{total} also makes an increase in the electric field of the tunnel oxide. Therefore, both the reliability of the tunnel oxide and the cell disturbance immunity are reduced by increasing the C_{FC}/C_{total} without optimizing the drain structure.

The drain structural dependence of the voltage scaling is also presented in figure 2. A decrease in the C_{FD}/C_{total} tends to reduce the V_D in the view of the capacitance coupling. However a decrease in the C_{FD}/C_{total} by reduction in the drain overlap area decreases the magnitude of the FN current. Therefore the drain voltage tends to increase with decrease in the drain overlap ratio in the view of the tunneling area. As a result, there exists the optimum drain overlap ratio for the V_D scaling as shown in figure 2. In other words, the drain voltage increases both for the large drain overlap cells due to the capacitance coupling effect and for the small drain overlap cells due to the FN tunneling area effect.

Moreover it can be seen that the suitable structure for low voltage operation depends on the C_{FC}/C_{total} in figure 2. At the low C_{FC}/C_{total} flash memories, the drain overlap effect plays an important role of determining the drain voltage as shown in figure 3(a). Therefore the decrease in the drain

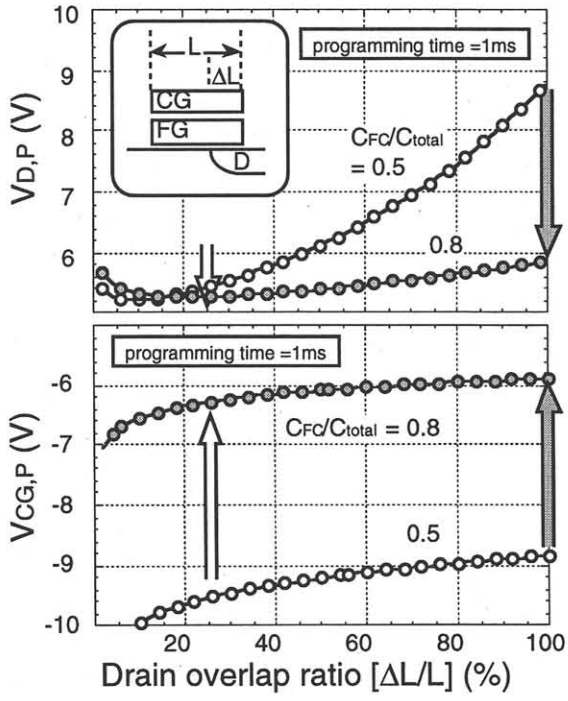

Fig. 2: Voltage Scaling of the V_D and the V_{CG} while programming, with a increase in the drain / gate overlap ratio as a function of the C_{FC}/C_{total}.

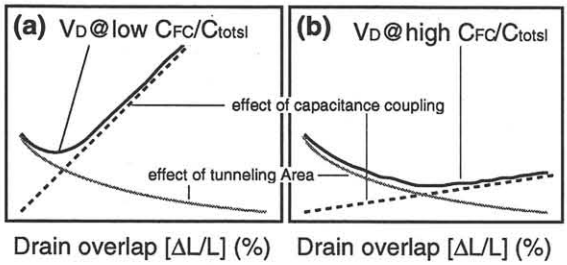

Fig. 3: The effect of the capacitance coupling and the tunneling area effect due to the drain overlap on the drain voltage at the high or the low CFC/Ctotal cell.

overlap, reducing the CFD/Ctotal, is effective for VD scaling. For an example, the drain voltage is only 5.1V by reducing the drain overlap ratio = 10% with the CFC/Ctotal = 0.5, compared with 8.8V at the drain overlap ratio = 100% as shown in figure 2. However, the control gate voltage exceeds -10V because the CFC/Ctotal is small. On the contrary, for the high CFC/Ctotal flash memories, the drain overlap effect decreases because the increment of the CFD/Ctotal with increase in the drain overlap area is suppressed by the high CFC/Ctotal. Consequently, the drain voltage is mainly controlled by the FN tunneling area effect as shown in figure 3(b). As a result, it is found that the high CFC/Ctotal can suppress the increase of the drain voltage even with increase of the drain overlap ratio. In our calculation, an increase of the drain voltage is less than 1V at the CFC/Ctotal = 0.8 even if the drain overlap ratio increases from 5% to 100% as shown in figure 2.

Supply voltage difference given by (|VD,prog. - VCG,prog.|) is also decreased for the high coupling cells. In figure 4, voltage difference is plotted as a function of the CFC/Ctotal or the drain overlap ratio. Right arrows in the figure 4 show the increase of supply voltage difference with increase of the drain overlap ratio. At the CFC/Ctotal = 0.8, an increase of the voltage difference is below 0.5V when the drain overlap ration increases from 20% to 100%, compared with the 2.5V difference at the CFC/Ctotal = 0.5.

Furthermore, the electric field across the tunnel oxide

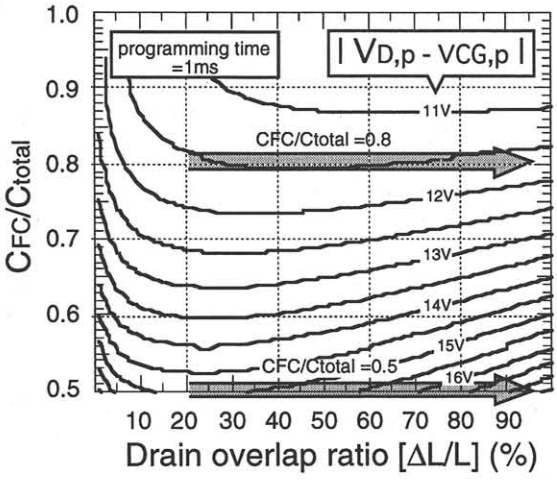

Fig.4: Voltage scalability for the coupling ratio and the drain overlap ratio. Right arrow shows the |VD-VCG| varification with increasing in the overlap ratio at coupling = 0.5 and 0.8.

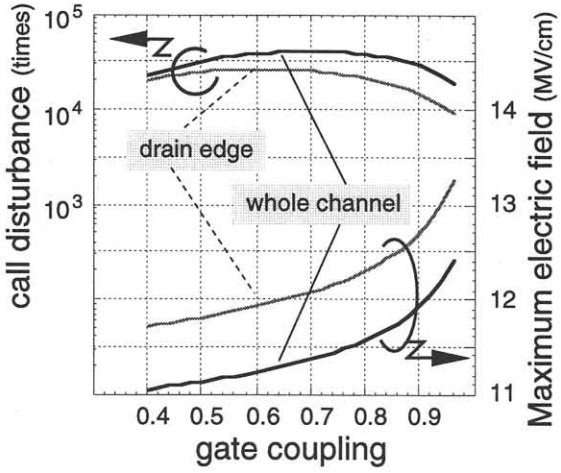

Fig.5: Comparison of the Emax of the tunnel oxide and the cell disturbance between the whole channel type (ΔL/L = 100%) and the drain edge type (ΔL/L = 25%).

and the cell disturbance is plotted in figure 5 at both cases of drain edge programming (ΔL/L=25%) and the whole channel programming (ΔL/L=100%). Based on the previous voltage scaling, the whole channel programming can decrease the maximum electric field. This is because the whole channel operation can keep the large amount of the FN gate current. Therefore when the programming time is kept constant we can reduce the tunnel oxide electric field which determines the FN current density. Moreover, the whole channel operation can also improve the cell disturbance. This is because reducing the maximum electric field makes reducing the tunnel oxide electric field at the disturbed cells. As a result the whole channel program is superior operation for low voltage operation, for high reliability of the tunnel oxide and for high disturbance immunity.

[Result]

A voltage scaling has been developed by the simple analytical equations to keep the program time constant with improving the tunnel oxide reliability and the cell disturbance immunity, while maintaining the constant tunnel oxide thickness. Using this scaling rule, the optimum structure for low voltage operation was investigated. As a result, it is demonstrated that the high coupling flash memories are suitable for low voltage operation. Moreover, the whole channel operation is useful for high reliability of the tunnel oxide and for the cell disturbance immunity. Furthermore, an increase in the drain voltage by using the whole channel programming can be suppressed by the high CFC/Ctotal. Therefore it is found that the high coupling with the whole channel programming is suitable for the high reliability and the low voltage operation of the FN type flash EEPROMs.

[References]

(1) J. D. Bude, et al., Tech. Dig. of IEDM 1995 p989
(2) H. Shirai, et al., Tech. Dig. of IEDM 1995 p653
(3) S. Ueno, et al., SSDM Extended Abstracts 1994 p518