A-2-8 A Proposition to Scale Down MNOS Nonvolatile Memory Devices

Yuji Yatsuda, Takaaki Hagiwara, Shin'ichi Minami, Ryuji Kondo,
Ken Uchida* and Kyotake Uchiumi*

Central Research Laboratory, Hitachi Ltd., Kokubunji, Tokyo 185 Japan
*Musashi Works, Hitachi Ltd., Kodaira Tokyo 187 Japan

The advent of Si-gate MNOS nonvolatile memories[1], featuring NMOS-compatible processes, is about to influence the usage of semiconductor memories.

This paper describes the fundamental considerations and experimental results of scaling down MNOS memory devices, focusing on the gate insulating film thicknesses. It also reports on the first 10-V programmable MNOS device.

FUNDAMENTAL CONSIDERATIONS In discussing the scaling down of MNOS devices, it is interesting to question whether ultra thin oxide (UTO) thickness (t_{ox}) should be scaled down, because the thickness strongly affects such MNOS characteristics of nonvolatile memory as (1) write-erase and (2) retention characteristics. As these characteristics are based on charge tunneling through the UTO, which has an exponential relationship to t_{ox}, the usual NMOS scale down theory, where all linear dimensions are reduced by a unitless scaling factor k[2], cannot be applied to t_{ox}, although it can be applied to silicon nitride (Si_3N_4) thickness (t_n).

On the basis of these characteristics, three cases of MNOS scale-down have been considered from a view point of total programming speed, as shown in Table I. In this table, case (1) features k^2 times the total programming speed (t_{pt}) with nearly constant programming speed per bit (t_p), retention time (t_r) and t_{ox}. Case (2) features constant t_{pt} with $1/k^2$ times t_p and t_r, and a slightly thinner t_{ox}. Case (3) has a much faster t_{pt} with 1/k times t_{ox} and much shorter t_p and t_r.

Case (1) or (2) should be suitable for obtaining more highly-integrated normal EEPROM's, and case (3) may be useful for special demands.

EXPERIMENTAL RESULTS AND DISCUSSIONS First, write and erase speed; t_w and t_e, of MNOS devices with oxide thicknesses (t_{ox}) of 2.02, 2.18, 2.29 and 2.36 nm, and with the same t_n of 50 nm were measured in order to determine whether the above-mentioned theoretical relationships are realized in Si-gate MNOS devices.

On the contrary, however, t_w is almost independent of t_{ox}, although t_e depends on t_{ox}, as expected. The results are shown in Fig.1. The difference between t_w and t_e dependences on t_{ox} is due to different carrier transport mechanisms; modified Fowler-Nordheim tunneling for writing and direct band-to-band tunneling for erasing. Therefore, it can be concluded that cases (2) and (3), featuring the faster t_p, are not useful for scaling down Si-gate MNOS memory devices.

Next, MNOS devices were scaled down according to case (1), i.e. keeping t_{ox} almost constant and 1/k times t_n. The relationships between t_w or t_e and V_w or V_e

were measured with different t_n as a parameter. The results are shown in Fig.2, which provides a means to determine t_n when t_w or t_e and V_w or V_e are given. As is seen, 10-V programming is possible if a t_n of 19.5 nm is chosen.

A t_n of 19.5 nm is almost the limit for down scaling because of charge centroid and the time constant for tunneling of charges trapped in Si_3N_4. Retention characteraretics of the device are shown in Fig.3. The measured charge decay rate is almost equal to that of unscaled devices.

Finally, if a programming speed faster than that of case (1) is needed, programming voltages larger than scaled down values should be used. For example, in the scaled down MNOS devices with 19.5 nm thick Si_3N_4, 12 V makes it possible to obtain a write speed faster than 1 μsec, as shown in Fig.2. Although fatigue phenomena appeared during an early write/erase cycle when high programming voltages were used, it was found that MONOS structure, in which the Si_3N_4 surface is oxidized, effectively eliminated the fatigue phenomena.

CONCLUSIONS Scaling down of MNOS devices was achieved by scaling down the Si_3N_4 film thickness while keeping the ultra thin oxide thickness constant. Thus, 10-V programmable MNOS devices are realized. High speed programmability was also cofirmed using a MONOS structure instead of conventional MNOS structures.

REFERENCES 1) T.Hagiwara et al., ISSCC '79, Digest of Technical Papers, P.50.
2) R.H.Dennard et al., IEEE J of S.S.C., Vol.SC-9, NO.5 (1974)

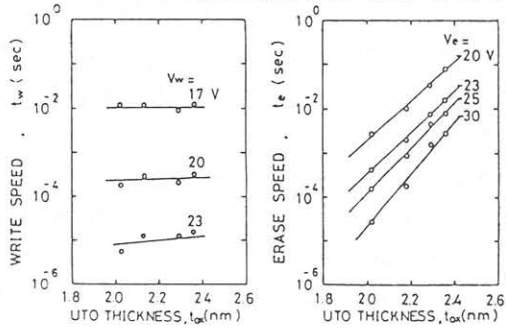

Fig.1 Write and erase speed dependences on ultra thin oxide film thickness

Table I Scaling factors for MNOS devices

Device parameter	Case (1)	Case (2)	Case (3)
UTO thickness, t_{ox}	1	<1 ($\ln(f_3(k))$)	1/k
Si_3N_4 thickness, t_n	1/k	1/k	1/k
Programming voltage, V_p	1/k	1/k	1/k
Total programming speed, t_{pt}	k^2	1	<<1 ($\exp(f_1(k))$)
Programming speed, t_p	-1*	$1/k^2$	<<1 ($\exp(f_1(k))$)
Retention time, t_r	-1*	$1/k^2$	<<1 ($\exp(f_2(k))$)

* This can be approximately obtained when $t_n/\varepsilon_n >> t_{ox}/\varepsilon_{ox}$.

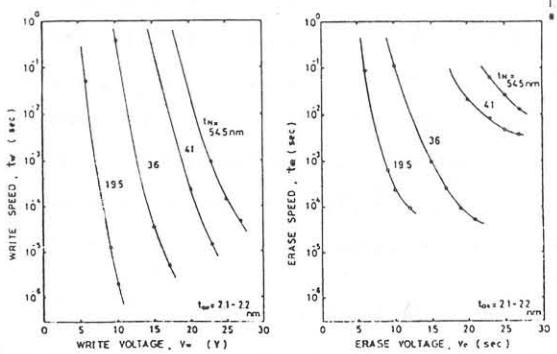

Fig.2 Write and erase speed dependences on programming voltage

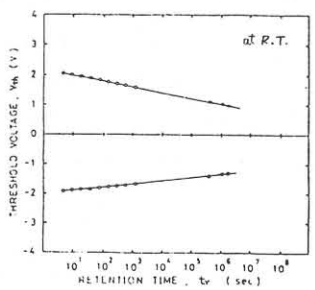

Fig.3 Retention characteristics of a scaled down MNOS device