High-Speed Multilevel NAND Flash Memory With Tight $V_{\rm th}$ Distribution Using an Engineered Potential Well and Forward-Bias Adjusted Programming

Gang Zhang, *Student Member, IEEE*, Zhe Wu, and Won Jong Yoo, *Senior Member, IEEE*

***Abstract*—This paper reports a high-speed multilevel-cell NAND Flash memory device using a Si–SiO_2–TiN–TiO_2–SiO_2–TaN (SOTTOT) engineered potential well (EW). The SOTTOT EW Flash memory device has very fast cell programming speed and good data retention. A 16-kbit NAND memory block using SOTTOT cells was programmed using a forward-bias-adjusted programming scheme, which enables bit adjustability during page programming to suppress the development of fast bits. The SOTTOT memory block shows fast programming speed ($\sim$40 μs/page), tight threshold voltage ($V_{\rm th}$) distribution ($\sim$0.22 V/level), and clear $V_{\rm th}$-level margins ($\sim$0.9 V) for the eight-level programming. The SOTTOT memory block also shows good resistance against pass/read disturbances as well as good ten-year data retention at an ambient temperature of 75 °C throughout 10^5 programming/erasing cycling.**

***Index Terms*—Engineered potential well (EW), forward-bias adjusted programming (FBAP), multilevel cell (MLC), NAND Flash memory.**

I. Introduction

MULTILEVEL-CELL (MLC) NAND Flash memory devices unify promising scalability and condensed data density, providing a favorable approach for mass data storage [1]–[7]. MLC storage programs a selected cell in a memory array to any n-value (with $n > 2$) different threshold voltages $V_{\rm th}$ so that each cell stores $b = \log_2 n$ bits of digital information. As a result, the data density is condensed, despite the device dimension, and the cost per bit is reduced for any lithographic technology generation [3]. In order to accomplish MLC storage for NAND memory arrays, various write-and-verify (WAV) schemes have been developed to pursue the precise programming of multiple $V_{\rm th}$ levels [4]–[6]. Wear leveling has been developed to improve the programming/erasing (P/E) endurance [7], and error-correcting techniques have been developed to improve the long-term data retention [7]–[9]. On the other hand, MLC NAND Flash memory devices using floating-gate and polysilicon–oxide–nitride–oxide–silicon (SONOS) memory cells are underdeveloped in terms of the programming speed due to WAV cycling. For example, it takes 200–300 μs to program a single-level cell, whereas it can take 600–900 μs to program an MLC [7]. Moreover, MLC NAND Flash memory devices may suffer from page programming disturbances [5], [6]. For example, a random page of memory cells are programmed simultaneously using the identical word-line (WL) bias condition, as shown in Fig. 1(a). Nevertheless, the programmed threshold voltages $V_{\rm th}$ read from all the bit lines (BL) must present a certain spread due to various (processing, environmental, etc.) disturbances, as shown in Fig. 1(b). The underprogrammed and overprogrammed cells are known usually as tail bits and fast bits, respectively. If WAV schemes complete the programming after all tail bits are programmed, as shown in Fig. 1(b), the fast bits may spread so far that they disturb the precise control of multiple $V_{\rm th}$ levels. This can be regarded as a type of programming disturbance. Such programming disturbances may be tolerable for single-level-cell NAND Flash memory devices, but it is critical for MLC NAND Flash memory devices, particularly for high density, e.g., 8- and 16-level MLC storage. In this regard, memory cells with faster programming speed are desired to speed up the WAV cycle, and programming schemes that suppress the programming disturbance are needed to improve the MLC $V_{\rm th}$ control.

In this paper, memory devices using a Si–SiO_2–TiN–TiO_2–SiO_2–TaN (SOTTOT) engineered potential well (EW) was proposed for a high-speed MLC NAND memory application. A SOTTOT memory device has a transitional boundary (i.e., the EW) between the tunnel barrier and the TiO_2 trapping layer. During P/E, the EW is bent by the gate electric field $E_{\rm ox}$, and direct tunneling (DT) of carriers via the shrunk tunnel barrier is enabled to accomplish rapid P/E. On the other hand, under retention mode, the tunnel barrier is extended due to charge trapping to suppress the discharge of trapped carriers [10], [11]. Therefore, the SOTTOT device has a significantly faster programming speed compared with the SONOS device,

Manuscript received December 21, 2010; revised May 30, 2011 and July 7, 2011. Date of publication August 22, 2011; date of current version September 21, 2011. This research was supported by the Basic Science Research Program of the National Research Foundation of Korea of the Ministry of Education, Science and Technology under Grant 2011-0006268 and Grant 2011-0010274. The review of this paper was arranged by Editor S. Deleonibus.

G. Zhang is with the Department of Nano Science and Technology, Samsung–SKKU Graphene Center, SKKU Advanced Institute of Nano Technology, Sungkyunkwan University, Suwon 440-746, Korea and also with the Electronic Materials Research Center, Korea Institute of Science and Technology, Seoul 136-791, Korea.

Z. Wu is with the Electronic Materials Research Center, Korea Institute of Science and Technology, Seoul 136-791, Korea and also with Samsung Electronics Company, Hwasung 445-701, Korea.

W. J. Yoo is with the Department of Nano Science and Technology, Samsung–SKKU Graphene Center, SKKU Advanced Institute of Nano Technology, Sungkyunkwan University, Suwon 440-746, Korea (e-mail: yoowj@skku.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2011.2162731

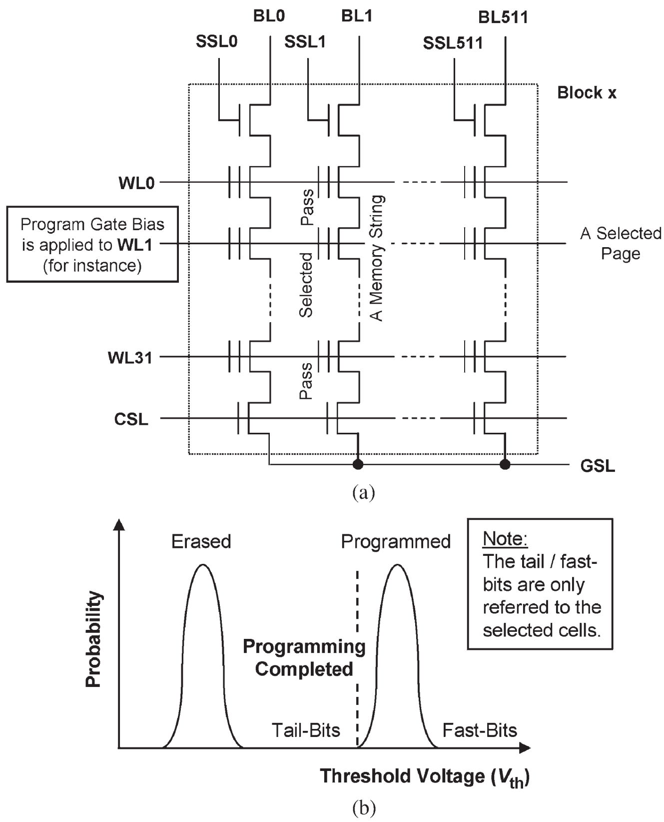


Fig. 1. (a) Equivalent circuit diagram of the experimental memory block. (b) Illustration of the tail bits, fast bits, and programming disturbances for NAND Flash memory devices.

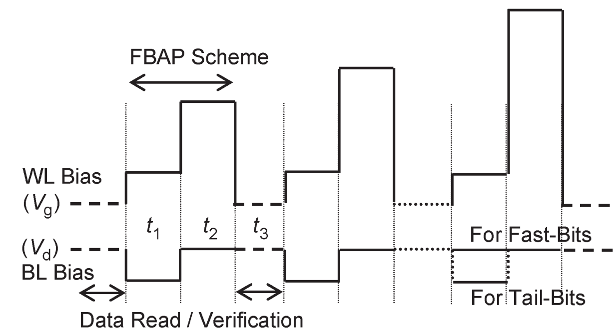


Fig. 2. Waveforms of WL (V_g) and BL (V_d) biases during FBAP. The forward-biased electron emission occurs during t_1, the preemitted electron injection occurs during t_2, and the V_{th} read/verification is performed during t_3.

and its data retention is good for long-term MLC data storage [11]. A 16-kbit NAND memory block using SOTTOT cells was programmed using the forward-bias-adjusted programming (FBAP) scheme, which enables bit adjustability for NAND memory blocks to suppress the development of fast bits during page programming. By modulating the gate bias V_g applied to the WL and the drain bias V_d applied to BLs for FBAP, 32 pages of SOTTOT cells were programmed to eight V_{th} levels with a very fast speed ($\sim$40 μs/page), tight V_{th} distribution ($\sim$0.22 V), and clear V_{th}-level margins ($\sim$0.9 V). The influence of pass/read disturbances during FBAP is insignificant, and the programmed V_{th} levels are expected to retain good ten-year data retention at an ambient temperature of 75 °C throughout 10^5 P/E cycling.

II. FBAP

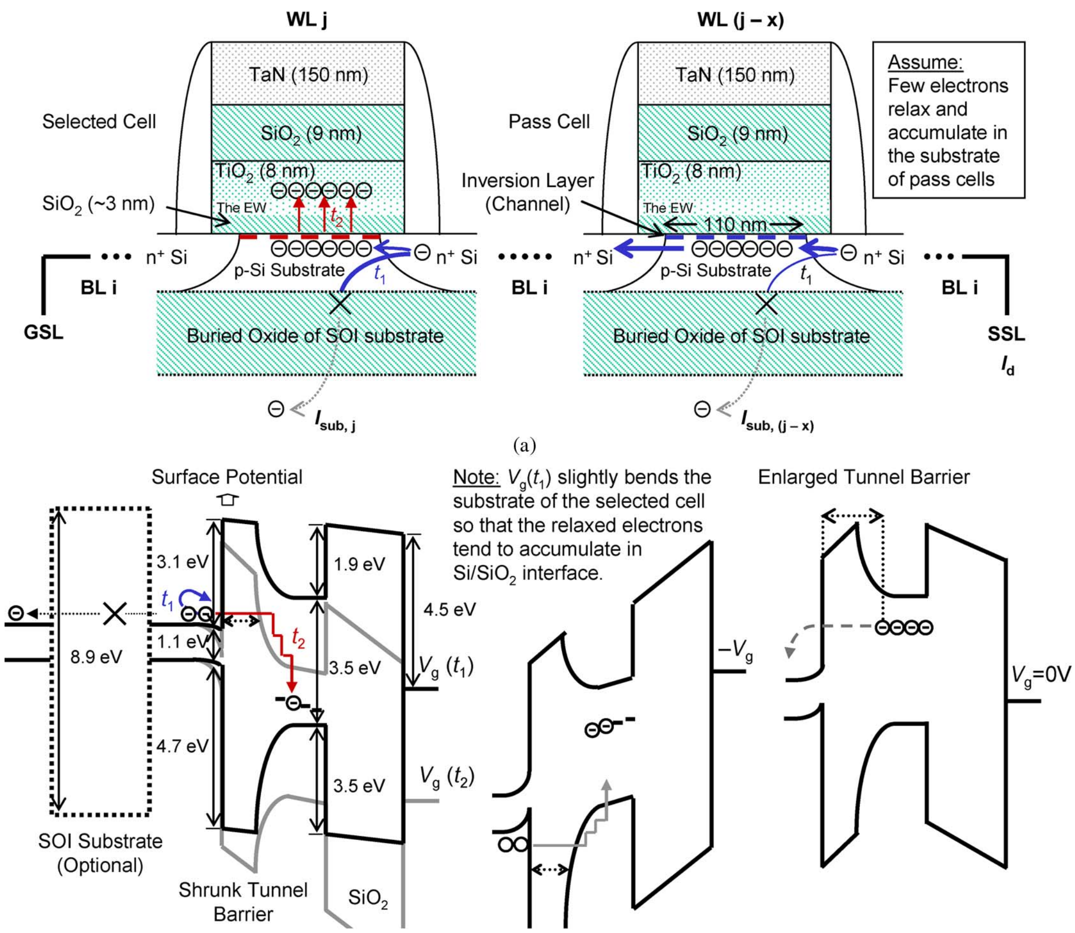
The FBAP scheme was developed based on the forward-bias-assisted electron injection (FBEI) method [12]. Similar to the FBEI method, the programming charges (electrons) are preemitted from the forward-biased p-n junction between the p-type substrate and the n^+-type drain before they are tunneled or injected through the tunnel barrier of a memory cell. On the other hand, DT and Fowler–Nordheim (F–N) tunneling of electrons induced only by $+V_g$ were employed for the FBAP scheme [10], [11], instead of hot-electron injection induced by both $+V_g$ and $+V_d$ for the FBEI method [12]. Fig. 2 shows the waveforms of the WL bias $+V_g$ and the BL bias $-V_d$ that performs FBAP. Before FBAP pulses are applied, the data states V_{th} of the selected cells [see Fig. 1(a)] are read to generate the appropriate programming parameters. During the period t_1 of an FBAP pulse, the confinement bias $+V_g < V_{\mathrm{th}}$ is applied to the selected WL, and $-V_d$ is applied to all the BLs. During t_2, the staircase programming bias $+V_g$ is applied to the selected WL while all BLs are grounded. During t_3, the read biases are applied to the selected WL and all BLs to read the programmed V_{th}. The periods of t_1, t_2, and t_3 need to be engineered according to the device structure. Meanwhile, the pass cells [see Fig. 1(a)] share the same BL bias with the selected cells simultaneously. On the other hand, the WLs of the pass cells are biased positively to turn on a continuous conducting channel for a memory string during t_1 and t_3, and the WLs of the pass cells are grounded during t_2 to avoid a disturbance to the selected page during programming. Table I lists the bias parameters applied to the selected cells and pass cells during an eight-level programming. Recall that the FBEI scheme was designed for single-level-cell nitride read only memory Flash memory programming, whereas the FBAP scheme was developed for MLC NAND Flash memory programming. The bias conditions as well as the applicability of FBEI and FBAP schemes are different.

Fig. 3 shows the mechanism of FBAP for a selected SOTTOT cell. During t_1, the p-n junction between the grounded p-type substrate and the negatively biased n^+ drain is forward biased; in which condition, the conduction electrons are emitted from the n^+ drain to the p-Si substrate, as shown in Fig. 3(a). $+V_g$ bends slightly the substrate so that the emitted electrons can accumulate near the Si/SiO_2 interface, as shown in the dark schematic in Fig. 3(b). These emitted electrons can be retained for $\sim$100 ns, which is the lifetime of free electrons in the p-Si substrate before their recombination with holes [12], [13]. When t_1 is designed properly, e.g., $<$ 100 ns, the emitted electrons can be direct tunneled readily to the trapping layer through the shrunk tunnel barrier when the large programming bias $+V_g$ is applied during t_2, as shown in the gray schematic in Fig. 3(b). After each FBAP pulse, the V_{th} of the selected memory cells are verified, and the next FBAP pulse with a step-up $+V_g$ bias is applied until the fastest bits reach the targeting V_{th} level. Subsequently, the BL bias $-V_d$ during t_1 is switched off for the faster bits by switching off their string select lines (SSLs), whereas the BL bias is still applied for the remaining

TABLE I
PROGRAMMING PARAMETERS FOR THE EIGHT-LEVEL PROGRAMMING

Programming Parameters	Targeting V_{th} (V)	WL Bias during t_1 (V)	BL Bias during t_1 (V)	WL Bias during t_2 (V)	BL Bias during t_2 (V)	WL Bias during t_3 (V)	BL Bias during t_3 (V)	Pulse Count (n+1)
Selected Cells		Adjustment Bias		Programming Bias		Read Bias		
111	≤1.4	-	-	–20	Grounded	1.4 - 1.5	0.15	
110	2.3	1	–1	13.0 + 0.05n	Grounded	2.3	0.15	~8
101	3.3	1	–1	13.5 + 0.05n	Grounded	3.3	0.15	~8
100	4.3	1	–1	14.0 + 0.05n	Grounded	4.3	0.15	~9
011	5.3	1	–1	14.5 + 0.05n	Grounded	5.3	0.15	~9
010	6.3	1	–1	15.0 + 0.05n	Grounded	6.3	0.15	~10
001	7.3	1	–1	17.0 + 0.05n	Grounded	7.3	0.15	~11
000	8.3	1	–1	19.0 + 0.05n	Grounded	8.3	0.15	~12
Pass Cells	-	Pass Bias		-		Read Bias		
All	Any	8.5	–1	Grounded	Grounded	8.5	0.15	-

n=integer 0, 1, 2, 3...



(b)

Fig. 3. (a) Illustration of the device structure of the SOTTOT cell and electron transfer via a pass cell and a selected cell during t_1 and t_2 of FBAP. The electrons tend to diffuse via the channel of the pass cell and to relax in the Si/SiO$_2$ interface of the selected cell (blue) during t_1. The relaxed electrons are injected to the trapping layer during t_2 (red). The relaxed electron density can be estimated by integrating $\Delta I_d = I_d - I_{\mathrm{sub},(j+j-x\ldots)}$. (b) Energy band diagrams of a SOTTOT cell during FBAP (where dark schematic shows electron emission and gray schematic shows electron injection)/erase/data retention. TEM image of the TiN/SiO$_2$ interface was obtained from [10]. Note that a constant confinement WL bias was applied in this paper.

cells until the slowest bits reach the targeting V_{th} level. Fig. 4 presents the operating logic of the FBAP scheme.

The inversion-layer electron density is approximately 1.49×10^{13} cm^{-2} in the SiO$_2$/Si interface of a MOS transistor of a substrate doping concentration of 7×10^{16} cm^{-3} under E_{ox} of 8 MV/cm, as shown by the dashed layer in Fig. 3(a). In contrast, the emitted electron density can be in the order of 10^{15} cm^{-2} near the SiO$_2$/Si interface when $V_g = 1$ V and $-V_d = -1$ V are applied for 100 ns to the same transistor, as shown by the symbolic electrons in Fig. 3(a). The emitted electron density was estimated from the transient characteristics of ΔI_d, which is defined as the difference in the drain current density I_d and the substrate current density I_{sub}. Fig. 5(a) shows the ΔI_d transient during a $-V_d$ pulse with the bias amplitude of -1 V and

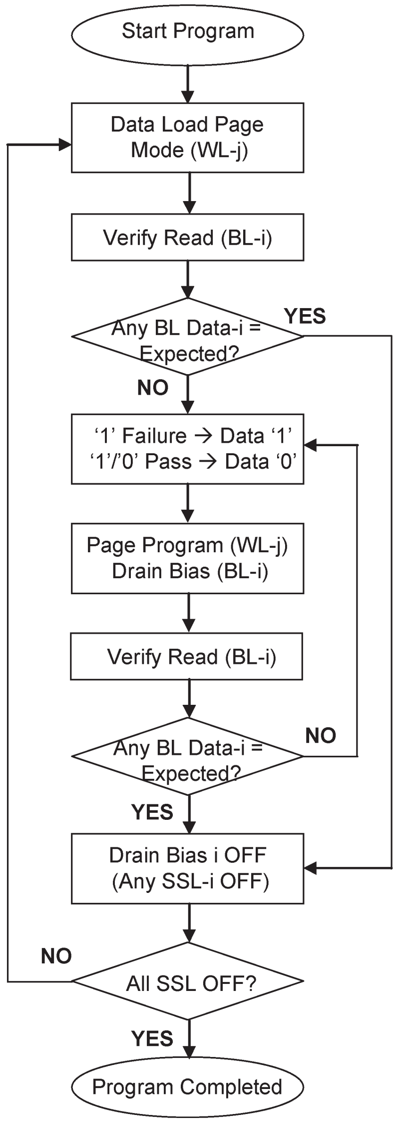


Fig. 4. Flowchart of operating logic for FBAP.

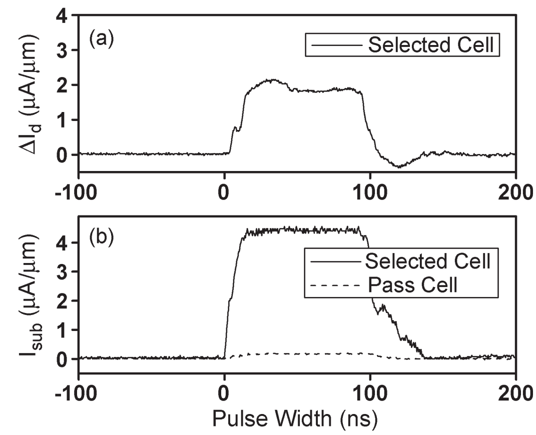


Fig. 5. (a) Transient characteristics of difference ΔI_d between drain and substrate current densities for a selected cell during the electron emission at t_1. (b) Transient characteristics of the substrate current densities $I_{\rm sub}$ for a selected cell and a pass cell during the electron emission at t_1.

the waveform of 2/100/2 ns applied to a MOS transistor built on a standard p-Si substrate (without the substrate buried oxide). ΔI_d is about 2 μA/μm during the 100-ns pulse. Since the emitted electrons are unlikely to diffuse beyond a localized area of $\sim$70 nm from the drain [13], the integration of ΔI_d along t may indicate the emitted electron density in the localized area of $\sim 1.7 \times 10^{15}$ cm^{-2}, which is approximately 10^2 times higher than the inversion-layer electron density. As a consequence, the programming speed of FBAP shall be enhanced effectively compared with that of F–N programming. On the other hand, the very high preemitted electron density increases certainly the surface potential of a memory cell, as shown in Fig. 3(b). The higher surface potential will overwhelm finally the confinement effect of $+V_g$ (during t_1) and drive the emitted electrons to the Si substrate. Therefore, the density of preemitted electrons must be less than the calculated results, and the programming speed of FBAP cannot be enhanced proportionally, compared with the ratio of the preemitted electron density versus the inversion-layer electron density. On the other hand, Fig. 5(b) presents the transient characteristics of $I_{\rm sub}$ for a selected cell and a pass cell. $I_{\rm sub}$ is $\sim$5 μA/μm for a selected cell, which is biased by $+V_g = 1$ V and $-V_d = -1$ V, whereas it is $\sim$0.16 μA/μm for a pass cell that is biased by $+V_g = 8.5$ V and $-V_d = -1$ V. As a conducting channel is missing in the selected cells, the emitted electrons tend to relax and to diffuse to the substrate to yield large $I_{\rm sub}$. In contrast, a conducting channel [i.e., the inversion layer in Fig. 3(a)] is formed in the pass cells during t_1. Therefore, the emitted electrons might tend to diffuse via the electrically less resistive conducting channel instead of via the substrate. As a result, the $I_{\rm sub}$ dissipation of pass cells is much smaller than that for the selected cells. In this manner, the emitted electrons are conducted from the BL to a selected cell via the memory string [see Fig. 1(a)]. Nevertheless, it is clear that $I_{\rm sub}$ dissipation is inevitable for the pass cells if they are fabricated on a standard Si substrate. This suggests that the conduction of $-V_d$ is limited via the memory string. Furthermore, $-V_d$ may induce significant power consumption when applied to all the selected and pass cells in a large memory block.

The emitted electrons can be confined well if the devices are fabricated on a fully depleted silicon-on-insulator (SOI) substrate. Using the floating-body effect of SOI devices, the emitted electrons can be confined for an extended period, i.e., it was reported that the emitted electrons can be retained for $10^{-6} - 1$ s in the floating body of a zero capacitor random access memory (Z-RAM) [14]. In this manner, the $+V_g$ during t_1 may be waived to simplify the FBAP scheme. The increase in the surface (body) potential of the floating body no longer drives the emitted electrons to the substrate for an SOI device, as shown in Fig. 3(b). Instead, the increased body potential might suppress the forward-biased I_d value rapidly to about zero. Therefore, the period of t_1 needs to be engineered carefully for the SOI devices. Nevertheless, it was assumed that the emitted electron density confined in the floating body may still be enhanced, and the efficiency of the FBAP scheme may be improved. Moreover, the accessibility of the pass cells shall be ensured because $I_{\rm sub}$ is cut off by the buried oxide layer, as illustrated in Fig. 3(a). Without $I_{\rm sub}$ dissipation, the power consumption will also be suppressed. In this paper, all the memory characteristics were obtained from the devices fabricated on a standard Si substrate rather than on a SOI substrate.

III. Memory P/E Performance

The memory transistors used in this paper have a SOTTOT (–/4/3/8/9/150 nm) gate stack [11] and a channel length of 110 nm. Fig. 3(a) and (b) shows the device structure and the energy band diagram of the SOTTOT EW device, respectively. The cross-sectional transmission electron microscopy (TEM) image of the EW was obtained in [10]. A thick (8 nm) trapping layer was used to enlarge the memory window $\Delta V_{\rm th}$, and a thick tunnel barrier (~3-nm-thick SiO_2 plus a ~3-nm-thick EW) was employed to improve data retention. Equivalent oxide thickness (EOT) of the memory cells is approximately 15 nm. The memory cells were integrated in a 16-kbit experimental memory block, which had 32 WLs and 512 BLs, as shown in Fig. 1(a). Note that each BL has an independent SSL, whereas all BLs share the same ground select line (GSL) in the experimental memory block. FBAP/F–N methods were used to perform P/E, and WAV cycling was applied to verify the programmed $V_{\rm th}$ levels. The WAV circuit was adopted in [4], and the SSL switch was achieved using a LABVIEW-program-controlled low-leakage switch mainframe. The negative BL bias was applied by an HP 4155A semiconductor parameter analyzer.

Fig. 6(a) shows the $V_{\rm th}$ window ($\Delta V_{\rm th}$) transients of SOTTOT cells programmed by FBAP and staircase F–N ($+V_g$ only) pulses. $\Delta V_{\rm th}$ was forward read at the read current I_d of 1 μA/μm [10]. The programming speed can be boosted effectively for devices programmed by FBAP, compared with those programmed using the F–N method. For example, it takes about ten FBAP pulses of $V_g = 1$ V and $-V_d = -1$ V during $t_1 = 100$ ns and V_g starting at 15 V ($E_{\rm ox} \approx 10$ MV/cm) with a step increase of 0.05 V during $t_2 = 100$ ns to obtain the $\Delta V_{\rm th}$ of 5 V, whereas it takes approximately 30 F–N pulses of V_g starting at 15 V with a step increase of 0.05 V for 100 ns to obtain the $\Delta V_{\rm th}$ of 5 V, as shown in Fig. 6(a). FBAP preemits a very high density ($< 10^{15}$ cm^{-2}) of electrons before programming. Therefore, the programming speed is faster than that of the F–N method [12], particularly when F–N programming rapidly approaches saturation under very low step-pulse biases. After removing the preemission of electrons, the FBAP pulses were converted to staircase F–N pulses, and its programming speed was drastically slower, as shown in Fig. 6(a). As the preemission of excess electrons is conducted by BL bias, which is manipulable during page programming, bit adjustability is enabled via an SSL switch for memory cells in a NAND array. This helps suppress the development of fast bits. For example, when the programmed $V_{\rm th}$ value of the fast bits reaches the targeting level, the forward BL bias is switched off for these cells for the subsequent WAV cycles (see Fig. 2). Consequently, the programmed $V_{\rm th}$ value saturates at the targeting level, as shown in Fig. 6(a). On the other hand, the BL bias is still applied for the rest of the cells so that the slower bits retain high-speed programming until the slowest bits reach the targeting $V_{\rm th}$ level. In this manner, fast bits are free from overprogramming,

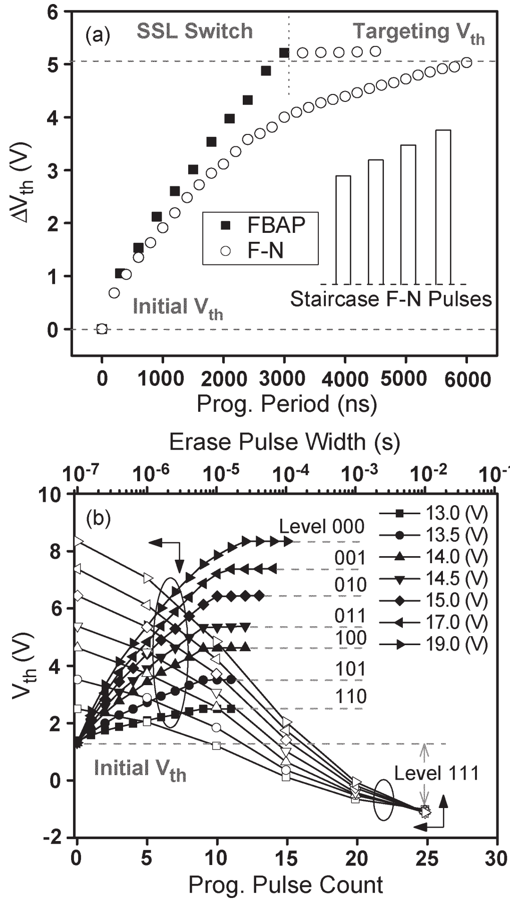


Fig. 6. (a) $\Delta V_{\rm th}$ transients of a SOTTOT cell programmed by FBAP and F–N methods. The initial $V_{\rm th}$ values are the same for FBAP and F–N programming. The staircase F–N pulses are illustrated inside. (b) $V_{\rm th}$ transients of a SOTTOT cell during the eight-level P/E.

whereas slower bits are programmed as usual. As a result, the $V_{\rm th}$ distribution is tightened effectively to allocate more data states within a finite $V_{\rm th}$ window.

MLC programming can be accomplished by modulating the amplitude of V_g (during t_2) and pulse count. Table I lists a typical scheme for an eight-level programming, and Fig. 6(b) presents the eight-level P/E characteristics of a random cell in the SOTTOT memory block. FBAP was applied for programming, and F–N method was applied for erasing. The energy band diagrams during erasing and data retention are reported in [11]. A sufficiently long (10 ms) erasing pulse of −20 V was used to erase all the programmed levels. The initial and (over-) erased $V_{\rm th}$ (≤ 1.4 V) values are regarded as the logic level 111. The programmed $V_{\rm th}$ levels with an average margin of ~0.9 V are regarded as the logic levels 110–000. With a period of 100 ns (t_3) for the programmed $V_{\rm th}$ read/verification, it takes ~40 μs to complete the programming of a SOTTOT memory page. This is significantly faster than that of F–N WAV programming (300–900 μs) for the floating-gate and SONOS memory devices [4]–[7]. This significantly faster page programming speed is due mainly to the very fast cell programming speed of the SOTTOT memory device, rather than the FBAP scheme, as compared with the other results in Table II. Furthermore, the large $V_{\rm th}$ window (~9 V) may also be due to FBAP, which injects high density

TABLE II
COMPARISON OF THIS PAPER TO THE OTHER REPORTED RESULTS

Technical features	This study	[1] *IEDM* 2010 s5p1, pp. 98-101	[5] *VLSI Tech.* 1995 pp. 129-130	[6] *VLSI Circuit* 1996 pp. 170-171	[18] *IEEE JSSC* 1995, vol. 30, pp. 1149-1156	[20] *IEDM* 2006 s2p1, pp. 19-22	[21] *IEEE JSSC* 2005, vol. 40, pp. 523-531
Device structure	SOTTOT	Floating-Gate	Floating-Gate	Floating-Gate	Floating-Gate	TANOS	Split tri-gated FG
Circuit	NAND	NAND	NAND	NAND	NAND	NAND	AG-AND
P/E mechanisms	FBAP / F-N	ISPP / F-N	ISPP / F-N	Intelligent ISPP / F-N	Page buffered ISPP / F-N	ISPP / F-N	CCIP / F-N
V_{th} verification	WAV	WAV	WAV	WAV	WAV	-	-
Bit-line adjustability	Yes	No	No	No	Yes	No	Split-gate control
Cell Prog. Speed	~10 μs	-	~200 μs	~200 μs	-	~100 μs	-
Typical Prog. speed	~40 μs/page	-	~300 μs/page	~600 μs/page	2.3 MB/s	-	10.3 MB/s
Typical V_{th} disper. (V)	~0.22	~0.7	~0.7	~0.5	<0.5	~1	~1
V_{th} level density	8-level	4/8-level	4-level	4-level	-	4-level	4-level
P/E endurance	$>10^5$ cycles	$>10^4$ cycles	-	-	-	$>10^4$ cycles	-
Data retention	~10-year	$>10^3$ hours@85°C	-	-	-	-	-

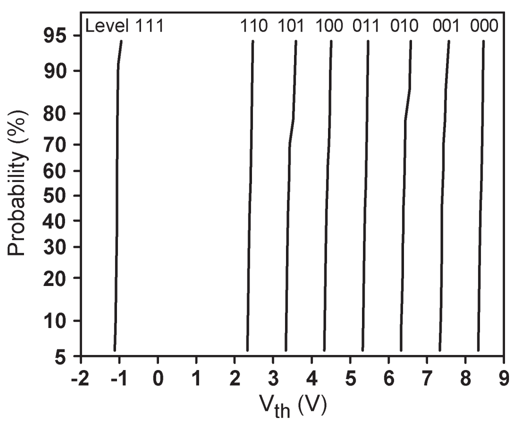


Fig. 7. Cumulative distribution of the P/E $V_{\rm th}$ levels for a random page in the memory block at the $10^{3,\rm th}$ P/E cycle.

of electrons to a thicker trapping layer [15] of the SOTTOT memory device. It is expected that the $\Delta V_{\rm th}$ of the SOTTOT memory device can be enlarged further by implanting ions to the TiO_2 trapping layer [16] to enable 16-level programming. In this paper, FBAP was applied for the SOTTOT memory device, which enables DT during programming, i.e., to demonstrate a very fast programming speed. The FBAP scheme shall present similar programming characteristics when applied for a band-engineered SONOS memory device, which also enhances DT during programming [17].

Fig. 7 shows the cumulative distribution of P/E $V_{\rm th}$ levels at the 10^3 P/E cycles for a random page in the memory block. The $V_{\rm th}$ dispersion was $\sim$0.22 V. This suggests that the fast bits are suppressed by the bit adjustability of FBAP. In contrast, the fast bits may lead to the typical $V_{\rm th}$ dispersion of $\sim$0.5–1 V for memory devices programmed by conventional WAV programming [5], [6]. On the other hand, the fine cumulative distribution of $V_{\rm th}$ levels suggests that the forward BL bias has been well conducted through no less than 31 pass cells. The SOI substrate can extend the conduction of forward BL bias effectively.

An incremental-step-pulse programming (ISPP) scheme also enables bit adjustability during page programming to tighten the $V_{\rm th}$ distribution [18]. Meanwhile, the ISPP may show linear programming transients ($\Delta V_{\rm th} \propto \Delta V_g$), which ease the programmed $V_{\rm th}$ control [19]. FBAP differs from ISPP in several aspects. FBAP switches the forward BL bias on/off before each programming cycle for tail/fast bits to enhance/suppress their programming speed, whereas ISPP applies the inhibition BL bias during each programming cycle for fast bits to suppress their programming speed. This means that FBAP is effective for both tail bits and fast bits, whereas ISPP is only effective for fast bits. Owing to the first aspect, FBAP may be more effective in tightening the $V_{\rm th}$ dispersion, as partially evidenced by the comparison in Table II. FBAP switches off the forward BL bias for more and more fast-programmed strings during WAV, whereas ISPP applies the inhibition BL bias to more and more fast-programmed strings. Therefore, FBAP must be less power consuming than ISPP. As a matter of fact, the power consumption of ISPP may become a critical concern when the inhibition BL bias ($>$ 3.3 V) is applied to a large number of strings, i.e., $> 10^3$, that are under the programming/pass WL biases ($V_g \gg V_{\rm th}$). In contrast, FBAP does not apply the WL bias ($> V_{\rm th}$) and the BL bias simultaneously. Nevertheless, FBAP and ISPP might be applied together (during t_2) to tighten the $V_{\rm th}$ distribution more aggressively.

IV. MEMORY RELIABILITY

The FBAP scheme well suppresses the programming disturbance for NAND Flash memory blocks, as shown in Figs. 6 and 7. On the other hand, due to the large WL and BL biases applied to the pass cells during t_1 and t_3, the pass disturbance and the read disturbance to the pass cells may become the major concerns for FBAP. In this paper, the influences of the pass and read disturbances were tested for the experimental SOTTOT memory block. Fig. 8(a) shows the eight-level $V_{\rm th}$ transients of a pass cell, which has undergone 10^3 P/E cycles, stressed by the pass bias (i.e., a WL bias of 8.5 V and a BL bias of -1 V) at an ambient temperature of 75 °C for 10^3 s to estimate the impact of a pass disturbance. The programmed $V_{\rm th}$ levels 000–110 were found to be insensitive to a pass disturbance, and the $V_{\rm th}$ variation was less than 0.1 V at these levels. On the other hand, the overerased level 111 is affected significantly by the pass disturbance. Level 111 is increased to $\sim$1.4 V after being stressed by the pass bias for 10^3 s. Nevertheless, level 111 is still close to its targeted $V_{\rm th}$ level ($\leq$ 1.4 V), and it might

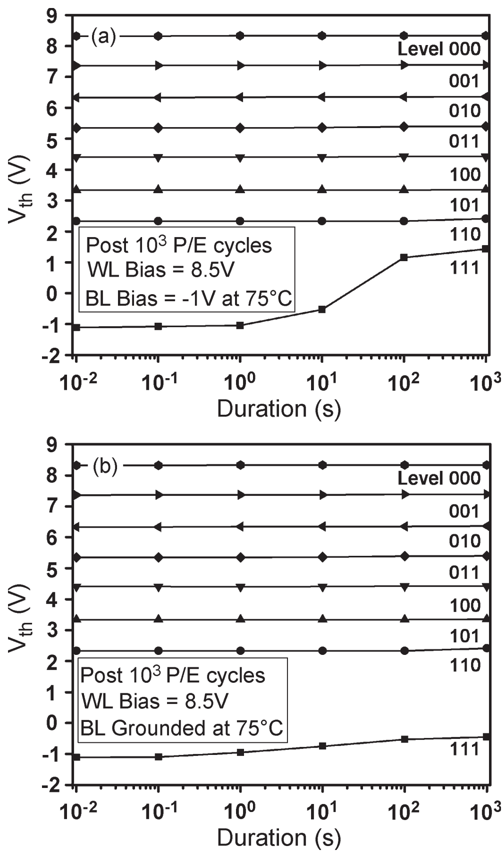


Fig. 8. (a) $V_{\rm th}$ transients of a SOTTOT cell (post-10^3 P/E cycles) stressed by $V_g = 8.5$ V and $V_d = -1$ V for 10^3 s at an ambient temperature of 75 °C. (b) $V_{\rm th}$ transients of the same SOTTOT cell stressed by $V_g = 8.5$ V for 10^3 s at an ambient temperature of 75 °C.

saturate at the targeting $V_{\rm th}$ level. In fact, the P/E endurance of the memory cell is required to be $10^5 - 10^6$ P/E cycles [7], which is equivalent to only $2 \times 10^{-2} - 2 \times 10^{-1}$ s. Therefore, the pass disturbance is tolerable for FBAP when it is applied to a SOTTOT memory block. Fig. 8(b) shows the eight-level $V_{\rm th}$ transients of the same SOTTOT cell stressed by the read bias (WL bias of 8.5 V) at 75 °C for 10^3 s to estimate the impact of the read disturbance. The programmed $V_{\rm th}$ levels 000–110 are immune to a read disturbance, and the overerased $V_{\rm th}$ level 111 is increased only slightly due to a read disturbance. A sufficient margin (> 2 V) remains between level 111 and the targeting $V_{\rm th}$ level after being stressed for 10^3 s (equivalent to 10^{10} read cycles). Therefore, FBAP is insensitive to the read disturbance. In this manner, a clear margin can be obtained between the programming bias (> 13 V) and the pass/read bias (≤ 8.5 V) for FBAP.

The electric field across the tunnel barrier $E_{\rm ox}$ can be estimated by $(V_g - V_{\rm FB})/\mathrm{EOT}$, where $V_{\rm FB}$ is the flatband voltage. When $V_{\rm th} > 1.4$ V and $V_g = 8.5$ V are applied, $E_{\rm ox} <$ 5 MV/cm, which is very small to disturb the programmed $V_{\rm th}$ levels of the tested SOTTOT cell [11]. On the other hand, when the tested SOTTOT cell is overerased, $E_{\rm ox}$ will be larger than 5.5 MV/cm, and the tunneling of preemitted electrons and channel electrons becomes inevitable. This explains the performances of the SOTTOT cell during pass and read disturbances. Nevertheless, these tunneled electrons can be captured

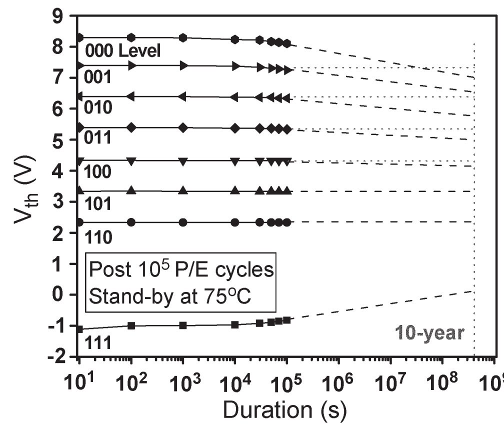


Fig. 9. $V_{\rm th}$ transients of a SOTTOT cell (post 10^5 P/E cycles) on standby at an ambient temperature of 75 °C for up to 10^5 s.

and trapped in the trapping layer to increase the potential of the trapping layer. As a consequence, the $V_{\rm th}$ of the tested cell increases and finally saturates at $\sim$1.4 V. The pass and read disturbances are suppressed subsequently.

Fig. 9 shows the eight-level $V_{\rm th}$ transients of a SOTTOT cell monitored for 105 s at the ambient temperature of 75 °C. The tested cell has undergone 10^5 P/E (between levels 111–000) cycles. The extrapolated 10-year data retention suggests that good data retention can be retained for the eight-level data storage at 75 °C. The SOTTOT device was reported to have an enlarged tunnel barrier and deep electron traps in the TiO_2 trapping layer to ensure the data retention [11]. Nevertheless, error-correcting techniques are still needed to correct the long-term error for $V_{\rm th}$ levels 000 and 001 [7]–[9]. The SOTTOT memory device is insensitive to P/E cycling-induced tunnel-barrier degradation. Therefore, it is expected that the SOTTOT memory device may tolerate more than 10^6 P/E cycles without degrading the memory performance [11]. Table II summarizes the memory performance of the SOTTOT memory device programmed by FBAP and the other reported results.

V. Conclusion

A SOTTOT EW Flash memory device programmed by FBAP was demonstrated for high-speed NAND MLC Flash memory applications. The SOTTOT memory device showed promising performance in high-speed MLC programming, very tight $V_{\rm th}$ distribution, a clear $V_{\rm th}$ margin, good resistance against pass/read disturbances, and good data retention.

References

[1] C.-H. Lee, S.-K. Sung, S.-H. Lee, S. Park, D.-H. Jang, J. Lee, S. Choi, E. Ahn, S. Kwon, H.-C. Baek, S.-S. Cho, J. Lim, J. Shin, J. Kim, K. Shin, K. Min, B. Yong Choi, S. J. Hwang, M. I. Kim, T.-H. Kim, M. Park, Y. Rah, J. Choi, K. Kim, J.-H. Choi, and T.-S. Jung, “A highly manufacturable integration technology for 27 nm multi-level NAND flash memory,” in *IEDM Tech. Dig.*, 2010, vol. s5p1, pp. 98–101.

[2] K. Prall and K. Parat, “25 nm 64GB MLC NAND technology and scaling challenges,” in *IEDM Tech. Dig.*, 2010, vol. s52, pp. 102–105.

[3] ITRS 2009. [Online]. Available: http://www.itrs.net/Links/2009ITRS/Home2009.htm

[4] T. Tanaka, M. Momodomi, Y. Iwata, Y. Tanakam, H. Oodaira, Y. Itoh, R. Shirota, K. Ohuchi, and F. Masuoka, “A 4-MBit NAND-EEPROM with

tight programmed V_{th} distribution," in *Proc. Symp. VLSI Circuits Dig.*, 1990, pp. 105–106.
[5] G. J. Hemink, T. Tanaka, T. Endoh, S. Aritome, and R. Shirota, "Fast and accurate programming method for multi-level NAND EEPROMs," in *VLSI Symp. Tech. Dig.*, 1995, pp. 129–130.
[6] Y.-J. Choi, K.-D. Suh, Y.-N. Koh, J.-W. Park, K.-J. Lee, Y.-J. Cho, and B.-H. Suh, "A high speed programming scheme for multi-level NAND Flash memory," in *Proc. VLSI Symp. Circuits Dig.*, 1996, pp. 170–171.
[7] J. Cooke, Flash memory technology direction. [Online]. Available: http://download.micron.com/pdf/presentations/events/WinHEC_Cooke.pdf
[8] S. Gregori, A. Cabrini, O. Khouri, and G. Torelli, "On-chip error correcting techniques for new-generation Flash memories," *Proc. IEEE*, vol. 91, no. 4, pp. 602–616, Apr. 2003.
[9] B. Chen, X. Zhang, and Z. F. Wang, "Error correction for multi-level NAND Flash memory using Reed-Solomon codes," in *Proc. IEEE Workshop SiPS*, 2008, pp. 94–99.
[10] G. Zhang, C. H. Ra, H.-M. Li, C. Yang, and W. J. Yoo, "Potential well engineering by partial oxidation of TiN for high-speed and low-voltage Flash memory with good 125 °C data retention and excellent endurance," in *IEDM Tech. Dig.*, 2009, vol. s34p5, pp. 835–838.
[11] G. Zhang, C. H. Ra, F.-M. Li, T.-Z. Shen, B.-K. Cheong, and W. J. Yoo, "Modified potential well formed by Si/SiO_2/TiN/TiO_2/SiO_2/TaN for flash memory application," *IEEE Trans. Electron Devices*, vol. 57, no. 11, pp. 2794–2800, Nov. 2010.
[12] S. S. Chung, Y. H. Tseng, C. S. Lai, Y. Y. Hsu, E. Ho, T. Chen, L. C. Peng, and C. H. Chu, "Novel ultra-low voltage and high-speed programming/erasing schemes for SONOS Flash memory with excellent data retention," in *IEDM Tech. Dig.*, 2007, vol. s17p4, pp. 457–460.
[13] G. Zhang and W. J. Yoo, "Pulse-agitated self-convergent programming for 4-bit per cell dual charge storage layer Flash memory," *Solid State Electron.*, vol. 54, no. 1, pp. 14–17, 2010.
[14] S. Okhonin, M. Nagoga, E. Carman, R. Beffa, and E. Faraoni, "New generation of Z-RAM," in *IEDM Tech. Dig.*, 2007, vol. s35p4, pp. 925–928.
[15] H.-W. You and W.-J. Cho, "Charge trapping properties of the HfO_2 layer with various thicknesses for charge trap Flash memory application," *Appl. Phys. Lett.*, vol. 96, no. 9, pp. 093506-1–093506-3, Mar. 2010.
[16] C. Y. Tsai, T. H. Lee, H. Wang, and A. Chin, "Highly-scaled 3.6 nm ENT trapping layer MONOS device with good retention and endurance," in *IEDM Tech. Dig.*, 2010, vol. s5p4, pp. 110–113.
[17] H.-T. Lue, S.-Y. Wang, E.-K. Lai, Y.-H. Shih, S.-C. Lai, L.-W. Yang, K.-C. Chen, J. Ku, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "BE-SONOS: A bandgap engineered SONOS with excellent performance and reliability," in *IEDM Tech. Dig.*, 2005, vol. s22p3, pp. 547–550.
[18] K.-D. Suh, B.-H. Suh, Y.-H. Lim, J.-K. Kim, Y.-J. Choi, Y.-N. Koh, S.-S. Lee, S.-C. Kwon, B.-S. Choi, J.-S. Yum, J.-H. Choi, J.-R. Kim, and H.-K. Lim, "A 3.3V 32 Mb NAND flash memory with increment -al step pulse programming scheme," *IEEE J. Solid-State Circuits*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995.
[19] H.-T. Lue, T.-H. Hsu, Y.-H. Hsiao, S.-C. Lai, E.-K. Lai, S.-P. Hong, M.-T. Wu, F. H. Hsu, N. Z. Lien, C.-P. Lu, S.-Y. Wang, J.-Y. Hsieh, L.-W. Yang, T. Yang, K.-C. Chen, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "Understanding STI edge fringing field effect on the scaling of charge-trapping (CT) NAND Flash and modeling of incremental step pulse programming (ISPP)," in *IEDM Tech. Dig.*, 2009, pp. 839–842.
[20] Y. W. Park, J. Choi, C. S. Kang, C. H. Lee, Y. C. Shin, B. H. Choi, J. H. Kim, S. H. Jeon, J. Sel, J. T. Park, K. H. Choi, T. H. Yoo, J. S. Sim, and K. N. Kim, "Highly manufacturable 32 Gb multi-level NAND Flash memory with 0.0098 μm^2 cell size using TANOS (Si-Oxide- Al_2O_3-TaN) cell technology," in *IEDM Tech Dig.*, 2006, vol. s2p1, pp. 19–22.
[21] H. Kurata, S. Saeki, T. Kobayashi, Y. Sasago, T. Arigane, K. Otsuga, and T. Kawahara, "Constant-charge-injection programming: A novel high-speed programming method for multilevel Flash memories," *IEEE J. Solid-State Circuits*, vol. 40, no. 2, pp. 523–531, Feb. 2005.

Gang Zhang (S'07) received the B.Eng. (with Hons.) degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2005, the M.Eng. and Ph.D. degrees in nanoelectronics from Sungkyunkwan University, Suwon, Korea, in 2008, and 2011, respectively.

From 2009 to 2011, he was a Research Assistant with the Electronic Materials Research Center, Korea Institute of Science and Technology, Seoul, Korea. He is currently with Samsung Electronics, Hwasung, Korea. His research interests include high-k dielectric materials and charge-trapping/phase-change nonvolatile memory.

Dr. Zhang was the recipient of the 2006 Best Student Paper Award at the Eighth International Conference on Solid-State and Integrated-Circuit Technology, Shanghai, China; the 2008 Gold Paper Award at the IEEE Student Paper Contest, Seoul, Korea; and the 2009 Chinese Government Award for Outstanding Self-Financed Student Abroad.

Zhe Wu was born in Yanji, Jilin, China, in 1981. He received the B.S. degree in materials, mechanical, and automation engineering from Yanbian University of Science and Technology, Jilin, in 2004, the M.S. degree in electronics from the University of Science and Technology, Seoul, Korea, in 2007, and the Ph.D. degree in material science and engineering from Korea Advanced Institute of Science Technology, Daejeon, Korea, in 2010.

He was a Research Assistant from 2005 to 2010 and a postdoctoral fellow from 2010 to 2011 with the Electronic Materials Research Center, Korea Institute of Science and Technology. He is currently a Senior Engineer with Samsung Electronics, Hwasung, Korea. His Ph.D. dissertation was focused on the characterization of Ge-doped SbTe-based chalcogenide materials for phase-change memory application. He also works on the development of parameter and dynamic random-access memory, and advanced Flash memory.

Dr. Wu was the recipient of The Korea Foundation for International Student Scholarship and Ilun Scholarship to support his M.S. and Ph.D. programs, respectively. He was a recipient of the Best Graduate Student medals during his B.S. and M.S. courses.

Won Jong Yoo (M'00–SM'02) received the B.S. and M.S. degrees from Seoul National University, Seoul, Korea, and the Ph.D. degree in the area of the plasma etching properties of Si and SiO_2 from Rensselaer Polytechnic Institute, Troy, NY, in 1993.

He was an Associate Professor with the National University of Singapore, where he conducted his research in the areas of silicon device and plasma processing. Since 2006, he has been with Sungkyunkwan University (SKKU) Advanced Institute of Nano-Technology, where he is currently a Professor. His main industrial experiences were research and development in the areas of semiconductor material/device processes with Samsung Semiconductor Research Center, Korea, and IBM Research Center, Yorktown Heights, NY. He is currently the Director of the Samsung-SKKU Graphene Center. He has authored or coauthored about 160 peer-reviewed journal and conference papers. The areas of his current research interests are the investigation of materials and electrical properties of graphene devices and nonvolatile memory devices using high-dielectric-constant materials, and the investigation of plasma etching processes of these devices.