A Multipage Cell Architecture for High-Speed Programming Multilevel NAND Flash Memories

Ken Takeuchi, Tomoharu Tanaka, and Toru Tanzawa

Abstract—**To realize low-cost, highly reliable, high-speed programming, and high-density multilevel flash memories, a multipage cell architecture has been proposed. This architecture enables both precise control of the V_{th} of a memory cell and fast programming without any area penalty. In the case of a four-level cell, a high programming speed of 236 μs/512 bytes or 2.2 Mbytes/s can be obtained, which is 2.3 times faster than the conventional method. A small die size can be achieved with the newly developed compact four-level column latch circuit. A preferential page select method has also been proposed so as to improve the data retention characteristics. The IC error rate can be decreased by as much as 43%, and a highly reliable operation can be realized.**

Index Terms—**Flash memory, high-speed programming, multilevel cell, NAND flash memory.**

I. Introduction

RECENTLY, a great deal of attention has been paid to multilevel flash memories because they drastically reduce the cost per bit. For example, the memory cell density can be doubled without a die size increase if the four levels of data can be stored in one memory cell. To realize multilevel flash memories, there are several difficulties which need to be solved. One requirement is that the programmed threshold voltage V_{th} of a memory cell needs to be controlled precisely. As the V_{th} distribution is larger, the maximum V_{th} is higher. The higher V_{th} needs a higher programming voltage, and adversely affects the data retention characteristics due to the larger electric field across the tunnel oxide during storage. Moreover, in a read operation, the read voltage V_{read} is applied to the control gate of unselected memory cells which are connected in series with the selected memory cell. Also, V_{read} must be higher than the maximum V_{th} of the memory cells so as to make the unselected memory cells act as transfer gates. Therefore, enlarged V_{th} results in a large V_{read} and degrades the read disturb characteristics.

On the other hand, although the multilevel cell scheme generally has the disadvantage of performance degradation, high-speed programming is still strongly demanded. The NAND flash memory uses Fowler–Nordheim tunneling for both the program and the erase to reduce power consumption, which allows page-based program operation and drastically increases program throughput. In designing a multilevel NAND flash memory, it is essential to decrease the program time per page because the target market is for solid-state mass storage applications, in which each program data, such as portrait data, have a large quantity and several pages must be continuously programmed. Furthermore, extremely slow programming causes an increase in test costs of both the memory chip and memory system.

Therefore, in developing multilevel NAND flash memories, the most important point is to control the programmed threshold voltage of memory cells precisely and shorten the program time per page. As a solution to this problem, a simultaneous multilevel program has been developed [1]. In this method, the threshold voltage during programming is controlled by the drain voltage under a constant control gate condition. Therefore, the memory cells on a wordline can be programmed to three different threshold voltages during the same programming period by changing the bitline voltage. Then, the three programmed levels could be verified cell by cell.

In this paper, it is shown that when the simultaneous multilevel program is used in NAND flash memories, the program speed will be drastically degraded. One reason is that the number of verify sequences is increased to three in a four-level cell system in comparison with just one sequence in a two-level cell system. The other is that the bitline voltage during programming is limited in the case of the self-boosted program inhibit voltage method [2], and as a result, three states actually cannot be programmed at the same speed.

To overcome this problem, a new architecture, a multipage cell architecture [3], is introduced. This architecture realizes both the precise V_{th} control and the fastest programming without any area penalty. A high programming speed of 236 μs/512 bytes or 2.2 Mbytes/s can be obtained, which is 2.3 times faster than the conventional method.

Since the target application of a multilevel flash memory is mass storage, the core circuit must be small so as to have a small die size. Especially, in NAND flash memories, the column latch has to be minimized because each bitline contains a column latch circuit for page program and page read operations [4], [5]. To meet this requirement, a compact, intelligent three-level column latch has been proposed [4], capable of successfully realizing precise V_{th} control, fast programming and small circuit area. Although several column latch circuits for a four-level cell have already been proposed [1], [5], [6], none of them can meet all of the above requirements. Some [1], [6] use multilevel sensing reference circuits, and can precisely control the threshold voltage of memory cells

Manuscript received September 29, 1997; revised February 25, 1998.

K. Takeuchi and T. Tanzawa are with the Microelectronics Engineering Laboratory, Toshiba Corporation, 1000-1, Kasama-cho, Sakae-ku, Yokohama 247, Japan (e-mail: ken.takeuchi@toshiba.co.jp).

T. Tanaka is with the Memory Division, Toshiba Corp., Yokohama, Japan.

Publisher Item Identifier S 0018-9200(98)05524-3.

0018–9200/98$10.00 © 1998

Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614

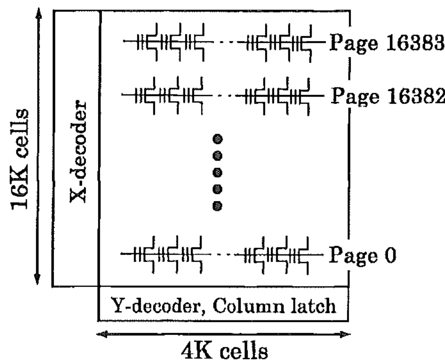


Fig. 1. A diagram of a 64-Mbit NAND flash memory [8].

using the cell-by-cell verify scheme, with the drawback of an area penalty. The other [5] minimizes the circuit area by eliminating reference circuits, but needs a longer program time because each state is programmed separately. In this paper, so as to achieve precise V_{th} control, fast programming, and small circuit area, a compact four-level column latch circuit was newly developed.

The data retention problem in the multilevel flash memory is much harder than the single-bit case, since the threshold voltage of a memory cell is higher and the electric field across the tunnel oxide is higher, and what is worse, the voltage differences between each level become smaller. To improve the data retention characteristics of a multilevel memory cell, a preferential page select method has also been proposed, where each memory cell is programmed preferentially to the lower V_{th} level. With the preferential page select method, the IC error rate decreases by as much as 43%, and operation at a highly reliable level can be realized.

In Section II, a basic array configuration of a NAND flash memory chip is reviewed. Next, the derivation of the program time is described in Section III. The concept of a multipage cell architecture is described in Section IV. The chip architecture of a proposed four-level cell is shown in Section V. The preferential page select method is given in Section VI. In Section VII, the compact four-level column latch is introduced, and key device operations are described. The performance improvements of the program and read operations are discussed, respectively, in Sections VIII and IX. Finally, the conclusion is provided in Section X.

II. Fixed Page Size

In NAND flash memories, the program operation is performed in page units so as to increase program throughput [7], [8]. Fig. 1 illustrates the 64-Mbit NAND flash memory [8]. A page is composed of memory cells which share the same control gate. The page size is 512 bytes. Even if the memory density is increased, this number will not change [2], [4], [5], [7], [8]. This is because the minimum data unit of a PC card is the same as the sector size of a hard disk. Since the page size cannot be increased, the program time per page has to be decreased.

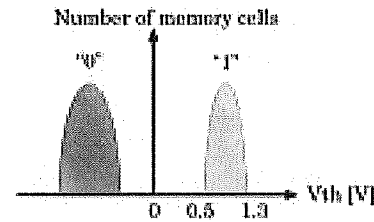


Fig. 2. Threshold voltage distribution of a two-level NAND cell.

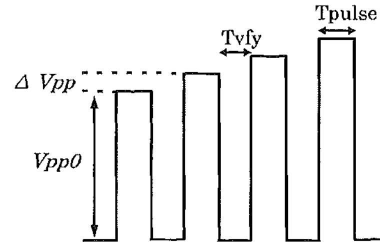


Fig. 3. Staircase programming pulses. A program verify operation is carried out after each pulse. ΔV_{pp} is the program voltage increment. V_{pp0} is the initial value of the program pulse. T_{vfy} is the verify read time. T_{pulse} is the program pulse width.

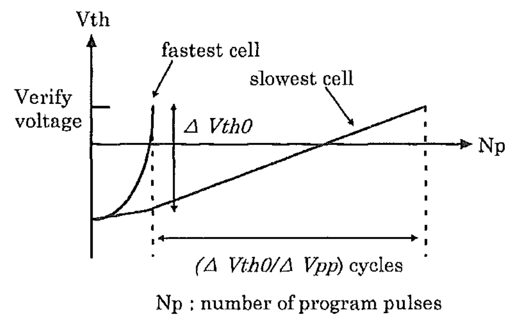


Fig. 4. Program characteristics of the fastest cell and slowest cell. At the first program pulse, the fastest cell is programmed sufficiently, and the V_{th} of the slowest cell is ΔV_{th0} lower than the verify voltage.

III. Derivation of the Program Time

Fig. 2 shows the V_{th} distribution of a two-level NAND cell. The "0" state is equal to the erased state. During the operation of the program, V_{pp} is applied to the control gate, while the channel of the cell is grounded. The staircase program pulse is shown in Fig. 3 [2], [9], where the verify steps are carried out after each pulse. In principle, the V_{th} distribution ΔV_{th} can be narrowed down to 0.5 V if ΔV_{pp} is 0.5 V [9]. In fact, the program voltage increment (ΔV_{pp}) and the control gate voltage during verify vary, respectively, by 0.1 V due to the V_{cc} variation. In addition, an array noise, such as a source line noise [10] or an interbitline capacitive coupling noise [11], increases ΔV_{th} by 0.1 V. As a result, ΔV_{th} increases to 0.8 V. Fig. 4 allows estimation of the number of program pulses N_p. At the first program pulse, the fastest cell is sufficiently programmed, and the V_{th} of the slowest cell is ΔV_{th0} lower than the verify voltage (0.5 V). Then, the V_{th} change of the slowest cell caused by each program pulse is constant with a value of ΔV_{pp}. Thus, $(1 + \Delta V_{th0}/V_{pp})$ cycles

Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614

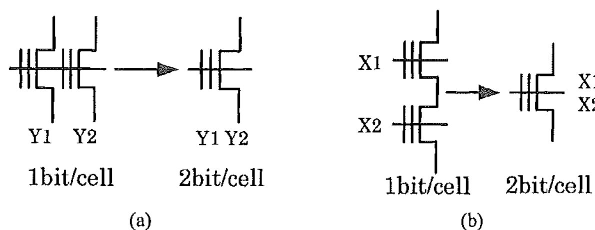


Fig. 5. (a) Conventional four-level cell. (b) Proposed four-level cell.

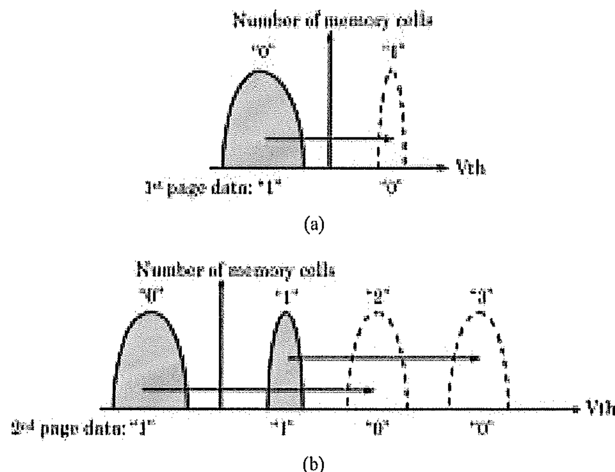


Fig. 6. Program operation of the proposed architecture. (a) First page program. (b) Second page program.

are necessary to program the slowest cell. Due to process variations and temperature variation, V_{pp} varies by δV_{pp} (0.5 V). Consequently, the number of program pulses N_p and the program time per page T_p are expressed as follows:

$$N_p = 1 + (\Delta V_{th0} + \delta V_{pp})/\Delta V_{pp} = 6$$

$$T_p = T_{load} + (T_{pulse} + T_{vfy}) \times N_p = 128\,\mu\text{s}.$$

ΔV_{th0} is 2 V. The duration of the data load T_{load} is 20 μs. The program pulse width T_{pulse} is 15 μs. The verify read time T_{vfy} is 3 μs.

IV. Concept of a Multipage Cell Architecture

Fig. 5 shows the conventional and the proposed four-level cell. The 2-bit data stored in a conventional cell correspond to two Y addresses ($Y1$, $Y2$). These 2 bit data belong to the same page. Also, three levels (the "1," "2," and "3" level) are programmed during the same operation [1], [5], [6]. On the other hand, the proposed cell contains two "pages." Here, a "page" means the group of memory cells that are programmed simultaneously. In other words, the 2-bit data of each cell correspond to two X addresses $X1$, $X2$, and the programming of $X1$ and that of $X2$ are performed during different operations. As shown in Fig. 6, at the first page program, the memory cell is programmed to the "1" state. After that, the "0" or "1" cell is, respectively, programmed to the "2" or "3" state during the second page program. Fig. 7 shows the V_{th} distribution. In the conventional method, the V_{th} distribution width of each programmed state is the same [1],

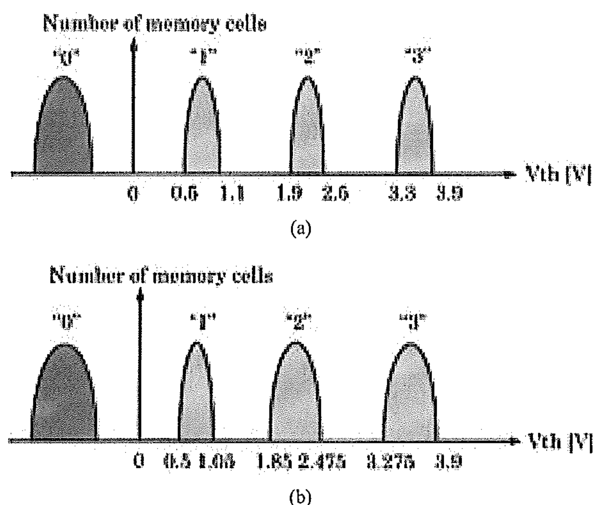


Fig. 7. Threshold voltage distribution of a four-level NAND cell. (a) Conventional. (b) Proposed.

[5], [6], as shown in Fig. 7(a). Conversely, the "1" program is controlled more precisely than the "2" or "3" program in the proposed method, as described below [Fig. 7(b)].

V. Chip Architecture

In NAND flash memories, the program operations are performed in page units to increase the program throughput. The page size is kept at 512 bytes in both two-level and four-level cells, as described in Section II. A schematic chip overview of the conventional [5] and the proposed 128-Mbit flash memory is shown in Fig. 8. Table I explains the core circuit configuration. In both the conventional [5] and the proposed four-level cell array, each column latch includes two latch circuits to store 2-bit data read out from or programmed into a four-level cell. Two adjacent bitlines share one column latch so as to reduce the column latch area. In both a conventional and a proposed four-level cell array, two bitlines sharing a column latch belong to different pages. And program or read is performed in quite different operations because two latch circuits are needed to program or read one memory cell connecting to one of the neighboring bitlines. In the conventional 128-Mbit cell array [5], 4K cells share the same control gate and 2K column latches are activated during a read or program operation. Although each column latch needs two latch circuits, the number of column latches is halved. Under the same page size condition, the total number of latch circuits in the conventional 128-Mbit four-level cell array is the same as the conventional 64-Mbit two-level cell array, and the circuit area does not increase. On the other hand, in the proposed four-level cell array, 8K cells share the same control gate so as to make the page size 512 bytes, and 4K column latches are included. Compared with the conventional 128-Mbit four-level cell array and the 64-Mbit two-level cell array under the same page size condition, there is no circuit area penalty because the number of X decoders is halved, whereas the circuit area of column latch is doubled, as seen in Table I.

Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614

TABLE I
CORE CIRCUIT CONFIGURATION OF THE 64-Mbit TWO-LEVEL NAND CELL ARRAY [8], THE CONVENTIONAL 128-Mbit FOUR-LEVEL NAND CELL ARRAY [5], AND THE PROPOSED FOUR-LEVEL NAND CELL ARRAY

	64Mb 2-level cell[8] (Fig.1)	Conventional 128Mb 4-level cell[5] (Fig.8(a))	Proposed 128Mb 4-level cell (Fig.8(b))
Page size	4kbit	4kbit	4kbit
Column latch circuit : Na	4k	2k	4k
Latch circuit per one column latch : Nb	1	2	2
Total latch circuit : Na×Nb	4k	4k	8k
X-address decoder (Block selector)	1k	1k	512

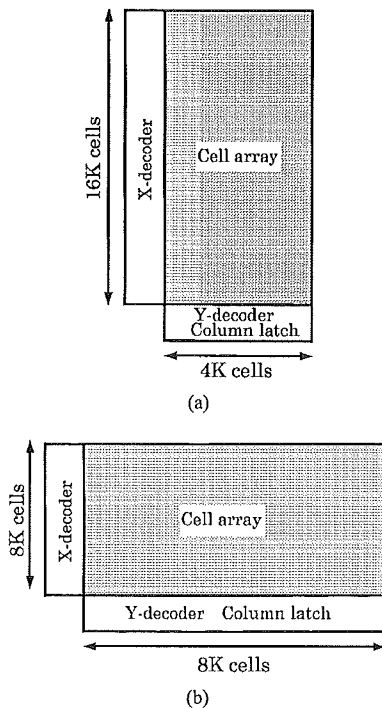


Fig. 8. Cell array organization of a 128-Mbit NAND flash memory. (a) Conventional [5]. (b) Proposed.

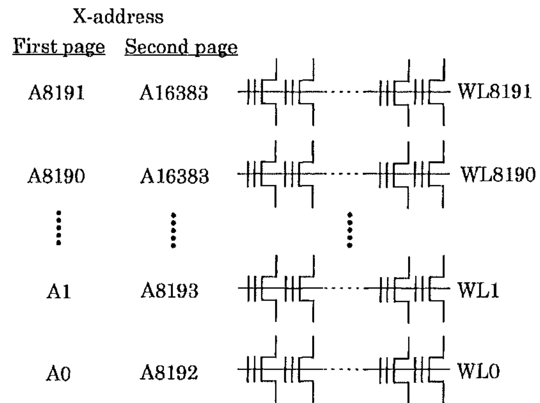


Fig. 9. Allocation of X addresses to memory cells in a proposed preferential page select method.

In NAND flash memories, the cell read current is as small as 1 μA. Therefore, the read access time is determined by the bitline capacitance, and the delay time of the wordline boosting has little effect on the access time. In the proposed array, as the bitline capacitance is halved, the (verify) read is accelerated.

VI. PREFERENTIAL PAGE SELECT METHOD

In a conventional NAND flash memory, the program cycle starts form the source-line side cell to the bitline side cell so as to prevent any interference between selected and unselected cells. Random page selection is not recommended during program operation. Such a restriction is not permitted in NOR flash memory, where the size of program data is small and random page access must be adopted. However, in NAND flash memory, the restriction of page selection does not slow down the program performance of the memory system because, in mass storage applications, each program data has a large quantity and many pages can be successively programmed. For example, in the case of a digital camera with a 400K-pixel CCD image sensor, the compressed image data are about 40 Kbytes, and as many as 80 pages can be successively programmed.

In the proposed multilevel NAND flash memory, a new page selection rule has been adopted. Each memory cell has a first page and a second page in it, as described above. During the program operation, the first page of each wordline must be selected first. In other words, the second page program before the first page program is inhibited. Such a restriction is acceptable in mass storage applications, as previously discussed.

Fig. 9 shows the X addresses of the proposed 128-Mbit four-level multipage cell array. The first half of all X addresses, that is, from "$A0$" to "$A8191$," corresponds to the first pages of the wordlines, from "$WL0$" to "$WL8191$." Also, the latter half of all X addresses, that is, from "$A8192$" to "$A16383$," corresponds to the second pages of the wordlines, from "$WL0$" to "$WL8191$." For example, the first X address "$A0$" corresponds to the first page of the wordline "$WL0$," and the next X address "$A1$" corresponds to the first page of the wordline "$WL1$." At first, all of the first pages, from "$WL0$" to "$WL8191$," are programmed. Next, the second page program is carried out to "$WL0$." In case the input data with a file size of 64 Mbits are stored in a 128-Mbit proposed cell array, all of the first pages are selected and all of the 64M memory cells are programmed to the "0" or "1" state, as shown in Fig. 10(b).

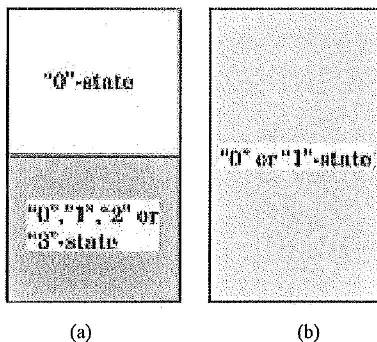


Fig. 10. 128-Mbit four-level memory cell array storing 64 Mbits of data. (a) In a conventional cell array, 32-Mbit memory cells are programmed to the "0," "1," "2," or "3" state, and the remaining 32M memory cells are still in the erased state. (b) In a proposed cell array, all of the 64M memory cells are in the "0" or "1" state.

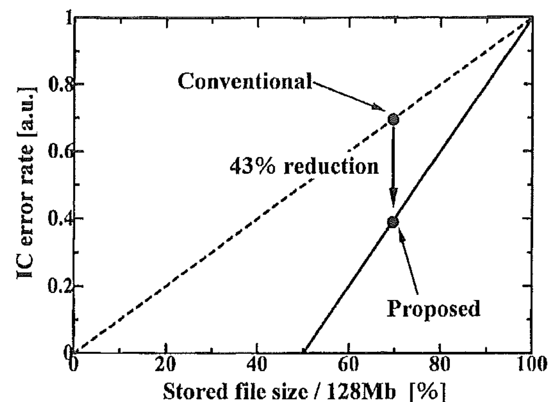


Fig. 11. Reliability improvement.

On the other hand, in the conventional method, 32M memory cells are programmed to the "0," "1," "2," or "3" state, and the rest are still in the erased states, as shown in Fig. 10(a). In the worst case, 32M memory cells are in the "3" states in the conventional method.

By using the proposed preferential page select method, the number of "3"-state cells decreases and data retention characteristics improve, as explained below. During shutdown (retention), the floating gate has a potential due to its stored charges. If the electric field across the tunnel oxide is high enough, electrons are ejected from the floating gate and V_{th} decreases. The leakage current during a retention is approximately described by the FN equation

$$I_{leak} = A \cdot E_{ox}^2 \cdot \exp(-B/E_{ox})$$

where A and B are characteristic constants and E_{ox} is the electric field across the tunnel oxide. E_{ox} during shutdown (retention) is expressed as

$$E_{ox} = \{C_{ono}/(C_{ox} + C_{ono})\}(V_{th} - V_{thi})/T_{ox}$$

where C_{ono}, C_{ox}, and T_{ox} are ONO capacitance, tunnel oxide capacitance, and tunnel oxide thickness. V_{thi} is a thermal equilibrium threshold voltage which is achieved by irradiating an ultraviolet light on a cell. In a NAND-type cell, V_{thi} is around 0 V to suppress a read disturb. Therefore, as V_{th} is higher, the electric field across the tunnel oxide E_{ox} is higher and the leakage current I_{leak} increases. In the case of a four-level cell, V_{th} of the "3" state is 2.8 V higher than that of the "1" state. A 2.8-V difference of V_{th} increases the leakage current by more than three orders of magnitude. As a result, the IC error rate during shutdown is determined by the failure rate of the "3" state to the "2" state. If the number of the "3"-state cells is decreased, the IC error rate also decreases. Fig. 11 shows the IC error rate. In this estimation, it is assumed that the failure rate of the "1" state to the "0" state is negligibly smaller than that of the "3" state to the "2" state. In the proposed method, the number of the "3"-state cells is drastically decreased. As a result, the data retention characteristics are definitely improved. In a solid-state file, only 70% of the total cells are counted as part of the capacity, and the rest is used as a work area that provides space for garbage collection [12]. In this case, the IC error rate decreases by as much as 43%, and a highly reliable operation can be realized.

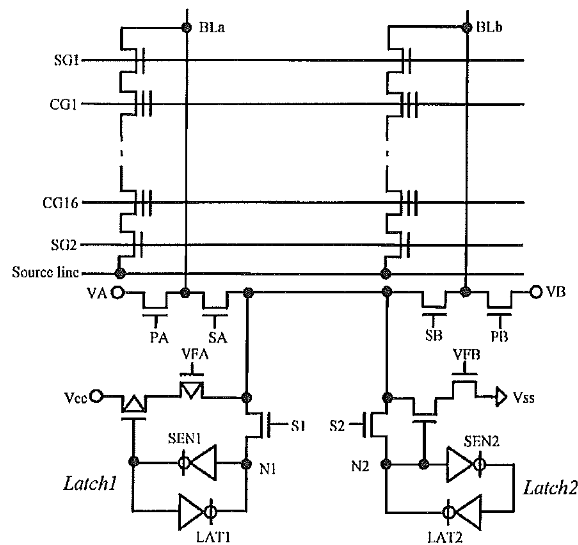


Fig. 12. Simplified schematic diagram of a proposed compact four-level column latch circuit.

VII. Operation

A. Compact Column Latch

The simplified circuit diagram of the proposed four-level compact column latch for a multipage cell is shown in Fig. 12. To obtain a small die size, the reference circuit is eliminated, and the number of transistors is minimized. Two bitlines share one column latch, and are alternately selected during the program and read operation. The data to be read out from or programmed into a memory cell are temporarily stored in a pair of flip-flop circuits, "Latch1" and "Latch2."

B. First Page Program Operation

Fig. 13(c) shows a flow chart of the program operation. The first page program is performed just like a two-level cell

Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614
5 of 11

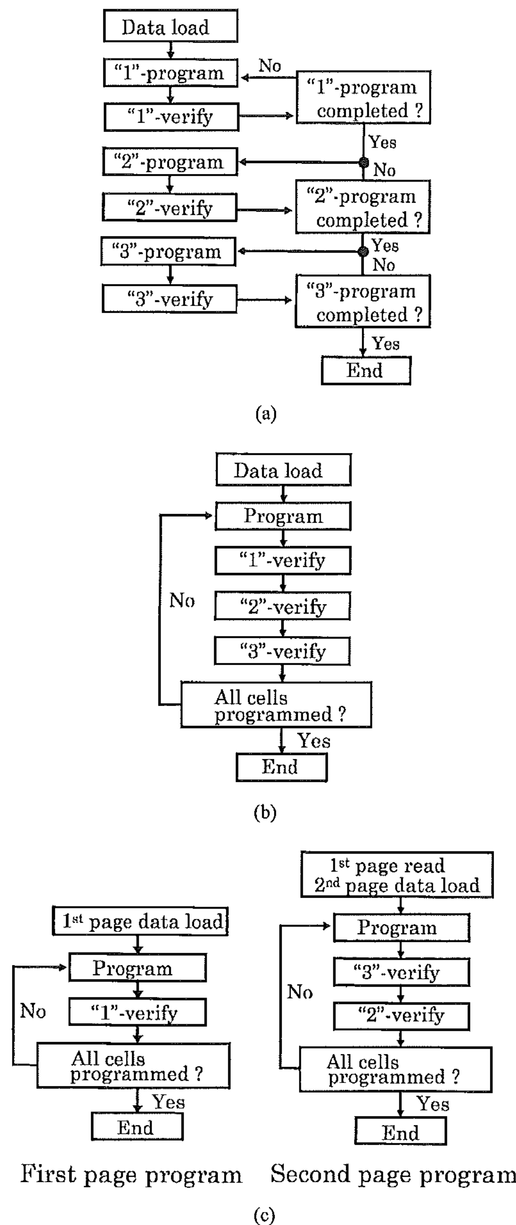


Fig. 13. Program algorithm of (a) the conventional four-level cell without simultaneous multilevel program [5], (b) the conventional four-level cell with simultaneous multilevel program [1], and (c) the proposed four-level cell.

[13]. The timing chart for the first page program operation can be seen in Fig. 14(a). The bitline *BLa* is selected, and the bitline *BLb* is unselected. The program data are input to "Latch1" as shown in Table II. Program pulses are applied to the selected control gate (CG). If the first page data are "1," the bitline voltage V_{ch} becomes V_{cc}. In this case, the select gate transistor connected to *SG*1 is turned off because *SG*1 is set to V_{cc}. As a result, the whole channel of the selected memory cell is coupled to wordline signals, and the memory cell remains in the erased state [2]. In the case in which the first page data are "0," V_{ch} is 0 V, and the cell is programmed to the "1" state. Program inhibit operation on an unselected bitline *BLb* is performed by supplying V_{cc} using the *PB* signal in Fig. 12. After the first page program operation, the first page verify read operation is carried out. Fig. 14(b) shows the timing chart of the first page verify read operation. The stored data in "Latch1" is modified such that programming is executed only when the memory cell is not sufficiently programmed. The program pulses are gradually raised from 18.3 V by ΔV_{pp}. As shown in Fig. 7(b), the V_{th} distribution ΔV_{th} of the "1" state is reduced down to 0.55 V by decreasing ΔV_{pp} to 0.25 V.

C. Second Page Program Operation

During the second page program, both "Latch1" and "Latch2" are used. The second page program data are input to "Latch1" through I/O, as can be seen in Fig. 15(b). During the data load, the inverse polarity of the first page data stored in the memory cell is also read out to "Latch2." The additional first page read does not slow down the program speed because "1" read is faster than the data load. Corresponding to the 2 bit data latched in the column latch, channel voltage V_{ch} is changed, as shown in Table III. Fig. 14(c) shows the timing diagrams of the second page program operation. In the case in which the second page data are "1," V_{ch} is V_{cc}, and the programming is inhibited. Thus, the memory cell stays at "0" or "1." If the column latch data *N*1 and *N*2 are, respectively, "0" and "1," V_{ch} is 0 V and the "1" cell is programmed to the "3" state. If both *N*1 and *N*2 are "0," V_{ch} is 1.425 V and the "0" cell is programmed to the "2" state. By raising the bitline voltage of the "2" program, the program speed of the "2" program is adjusted to be the same as the "3" program. After the second page program operation, the second page verify read operation is performed. The timing chart of the second page verify read operation is shown in Fig. 14(d). The second page verify read operation is composed of the "3" and "2" verify. During the "3" and "2" verify, the selected control gate is applied, respectively, to 3.275 and 1.85 V. In the case of a two-level cell [13] or a three-level cell [4], after the memory cell data are read out to the bitline, the bitline voltage is modified according to the 1 bit data stored in one flip-flop circuit. To realize the four-level cell-by-cell verify read operation, the key operation is that, after the data of the selected memory cell are read to the selected bitline, the bitline voltage is changed according to the 2-bit data in both "Latch1" and "Latch2" by pulsing *VFA* and *VFB*. As a result, when a memory cell is programmed to the target V_{th} of its respective state, the associated "Latch1" is automatically reset to inhibit further programming by providing V_{cc} to the selected bitline.

D. Read Operation

The random read of the first and second pages is performed during the same operation. One of four different states is identified by sequentially changing the wordline voltage [4], [5]. Therefore, the multiple reference circuit [1], [6] is eliminated, and a much smaller circuit area can be obtained. The timing diagram is shown in Fig. 14(e). The read operation is composed of three phases, and the wordline is applied to three different levels corresponding to three programmed

Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614
6 of 11

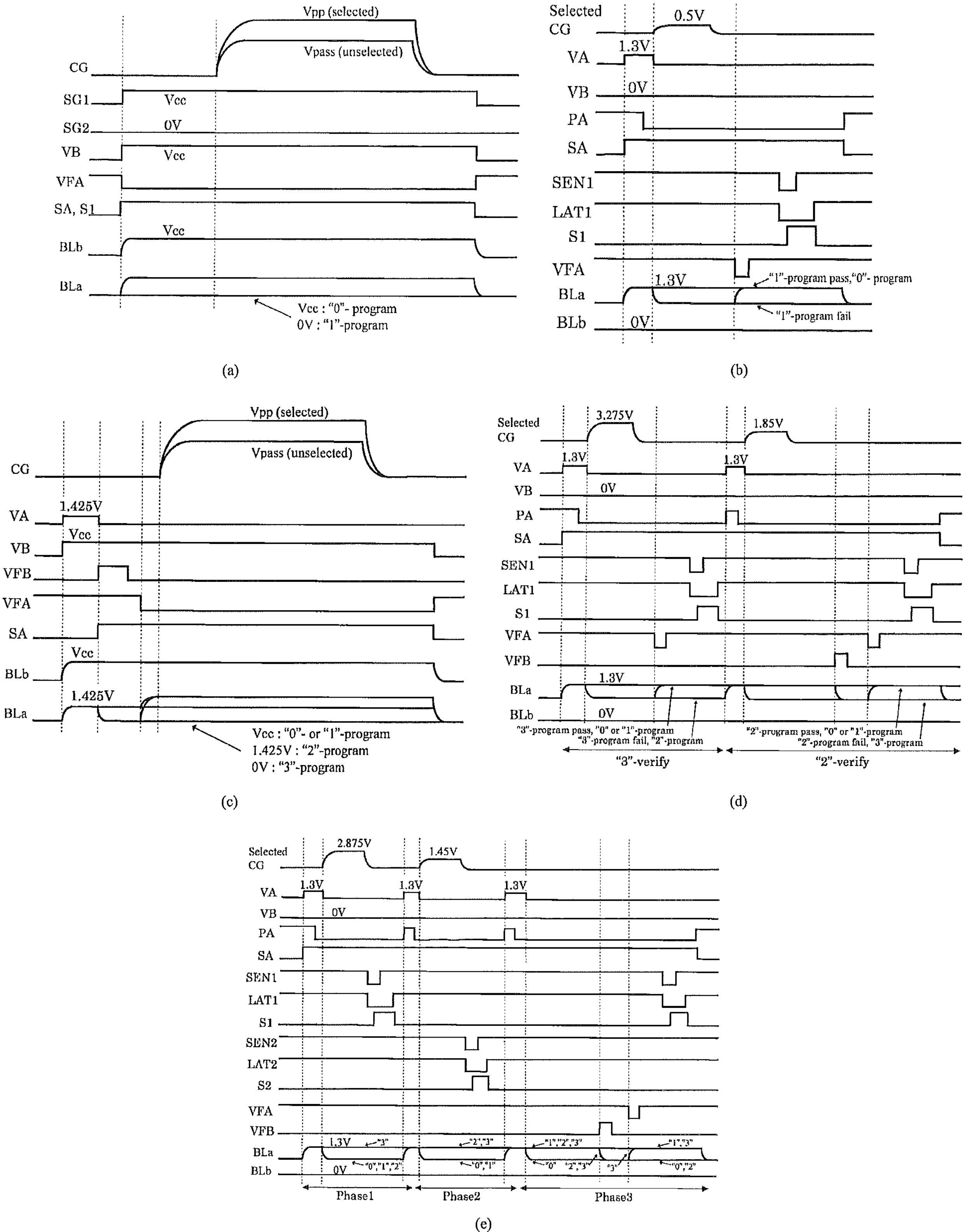


Fig. 14. Signal timing diagram. (a) First page program operation. (b) First page verify-read operation. (c) Second page program operation. (d) Second page verify-read operation. (e) Read operation.

Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614
7 of 11

TABLE II
RELATION BETWEEN PROGRAM DATA STORED IN A COLUMN LATCH CIRCUIT AND THE BITLINE VOLTAGE V_{ch}, DURING THE FIRST PAGE PROGRAM

	Memory cell state	
	"0"-state	"1"-state
1st page data N1	1	0
bitline voltage Vch	Vcc	0V

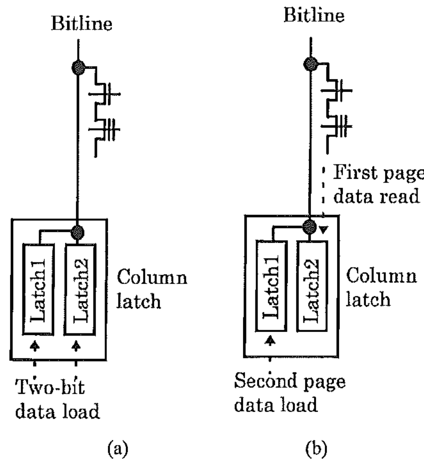


Fig. 15. Data load operation of (a) the conventional method and (b) the second page program of the proposed method.

TABLE III
RELATION BETWEEN PROGRAM DATA STORED IN A COLUMN LATCH CIRCUIT AND THE BITLINE VOLTAGE V_{ch}, DURING THE SECOND PAGE PROGRAM

	Memory cell state			
	"0"	"1"	"2"	"3"
2nd page data N1	1	1	0	0
1st page data N2	0	1	0	1
bitline voltage Vch	Vcc		1.425V	0V

states. During the first phase, the selected control gate is biased to 2.875 V, and the bitline data are latched in "Latch1." At the time of the second phase, the selected control gate is set to 1.45 V, and then the bitline data are latched in "Latch2." Finally, at the third phase, the selected control gate is applied to 0 V. After the memory cell data are read out to the selected bitline *BLa*, the bitline voltage is modified according to the stored data in both "Latch1" and "Latch2" by activating the *VFB* and *VFA* signals. Then, the bitline data are latched to "Latch1." The final data are shown in Table IV. The first and the second page data are latched, respectively, in "Latch1" and "Latch2."

VIII. PROGRAM TIME COMPARISON

A. Conventional Method Without Simultaneous MultiLevel Program [5]

Fig. 13(a) shows the program operation of the conventional method without a simultaneous multilevel program [5]. The program operation is composed of three program cycles. During the data load, 2 bit program data are input to the column latch circuit, as seen in Fig. 15(a). At the first program cycle, programming and program verify of the "1" state are repeated until the "1" program is completed. Next, at the second program cycle, the "2" program and "2" verify read are operated. Finally, at the third program cycle, the "3" program and "3" verify are performed. Each program operation is the same as the two-level cell, that is, the selected bitline is grounded and the program inhibit bitline is biased to V_{cc}. The program voltage increment ΔV_{pp} is 0.3 V to make ΔV_{th} 0.6 V. Therefore, the number of program pulses N_p during each program cycle is expressed as follows:

$$N_p = 1 + (\Delta V_{th0} + \delta V_{pp})/\Delta V_{pp} = 10.$$

As a result, the total program time T_p, which is the sum of the first, second, and third program cycles is as follows:

$$T_p = T_{load} + 3 \times (T_{pulse} + T_{vfy}) \times N_p = 695\,\mu\text{s}.$$

The verify read time T_{vfy} is prolonged to 7.5 μs because the maximum V_{th} is higher than the two-level cell, and the decreased voltage difference between the control gate and V_{th} reduces the cell read current. As seen in Fig. 17, in case the simultaneous multilevel program is not used, the program characteristics are seriously degraded because each state is programmed separately.

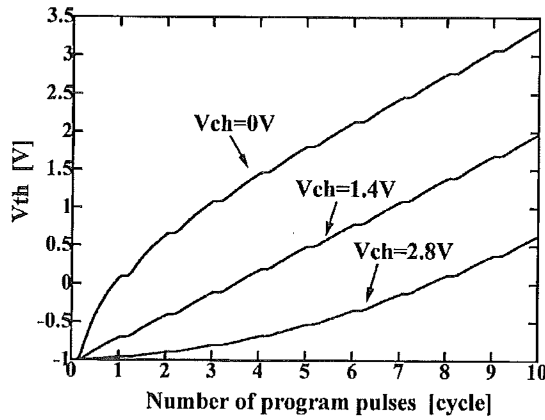
TABLE IV
READ-OUT DATA LATCHED IN A COLUMN LATCH CIRCUIT

	Memory cell state			
	"0"	"1"	"2"	"3"
1st page data N1	0	1	0	1
2nd page data N2	0	0	1	1

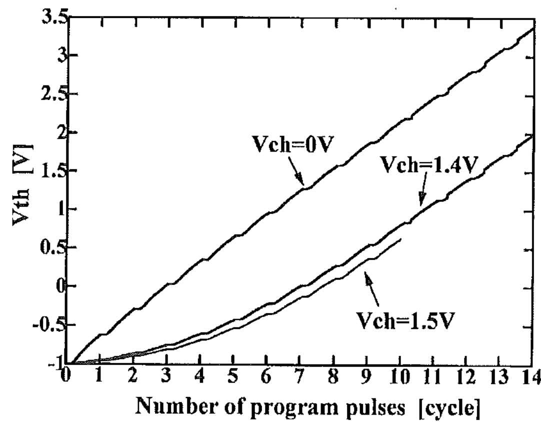
B. Conventional Method with Simultaneous Multilevel Program

Fig. 13(b) shows the program operation of the conventional method with a simultaneous multilevel program. During the data load, 2-bit program data are input to the column latch circuit, as seen in Fig. 15(a). During the program, the channel voltage of the memory cell V_{ch} is changed according to the stored data in the column latch. Then, three verify read sequences ("1," "2," "3" verify) are subsequently carried out, where the control gate is applied to verify voltages 0.5, 1.9, and 3.3 V, respectively. Ideally, V_{ch} is 0, 1.4, and 2.8 V for the "3," "2," and "1" program where 1.4 V (2.8 V) corresponds to the V_{th} difference between the "3" and "2" ("1"). The simulated write characteristics of the slowest cell are shown in Fig. 16(a). ΔV_{pp} is 0.3 V to make ΔV_{th} 0.6 V, as can be seen in Fig. 7(a). The three levels are programmed at the same speed. Including the V_{ch} variation δV_{ch}, the number of program pulses N_p and

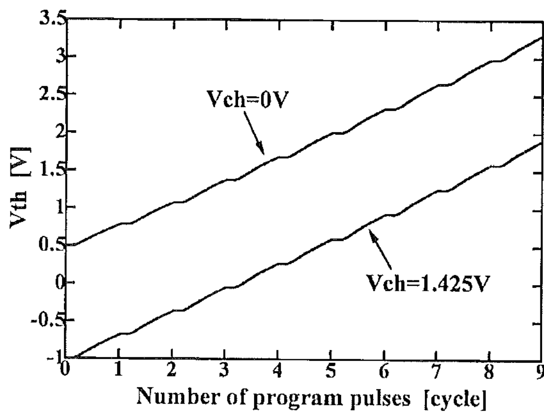
Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614



(a)



(b)



(c)

Fig. 16. Simulated program characteristics of the slowest cell. (a) Ideal case of the conventional method. V_{pp0} is 18.3 V, and ΔV_{pp} is 0.3 V. (b) Actual case of the conventional method. V_{pp0} is 17 V, and ΔV_{pp} is 0.3 V. (c) Proposed method. V_{pp0} is 18.3 V, and ΔV_{pp} is 0.325 V. The couple factor of the cell was 0.57, and the tunnel oxide thickness was 10 nm.

the program time per page T_p are as follows:

$$N_p = 1 + (\Delta V_{th0} + \delta V_{pp} + \delta V_{ch})/\Delta V_{pp} = 10$$
$$T_p = T_{load} + (T_{pulse} + 3 \times T_{vfy}) \times N_p = 395\,\mu\text{s}.$$

δV_{ch} is 0.1 V. The verify read time T_{vfy} is 7.5 μs. Compared with the two-level cell, the increased verify sequences, that occupy as much as 57% of the program time, seriously slow down the program speed, as shown in Fig. 17.

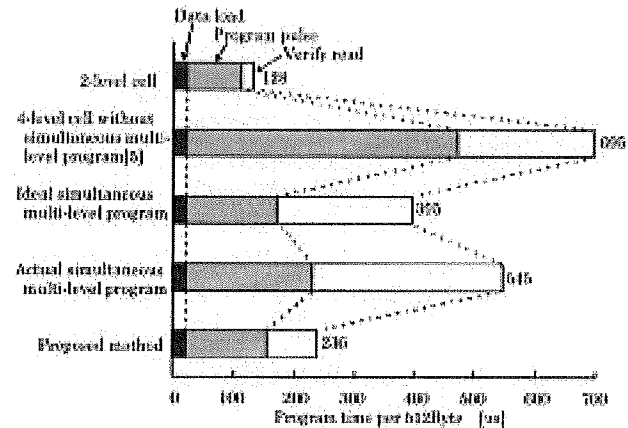


Fig. 17. Program time comparison.

Actually, in the case of the "0" program, the bitline and the select gate connected to the bitline are applied to V_{cc}, and the select gate transistor is turned off [2]. If the minimum V_{cc} is 2.8 V, V_{ch} must be lower than 1.5 V so as to turn on the select gate transistor during the "1," "2," or "3" program. Thus, V_{ch} is 0 V for the "3" program, 1.4 V for the "2" program, and 1.5 V for the "1" program. The initial value for V_{pp}, V_{pp0}, must be δV_{bl} lower than the ideal case, where δV_{bl} is the V_{ch} difference between the actual and ideal case during the "1" program. If this is not done, the fastest cell would be programmed higher than the "1" state. As a result, the "2" and "3" program become slower than the ideal case, as shown in Fig. 16(b), and the number of program pulses N_p definitely increases:

$$N_p = 1 + (\Delta V_{th0} + \delta V_{pp} + \delta V_{ch} + \delta V_{bl})/\Delta V_{pp} = 14$$
$$T_p = T_{load} + (T_{pulse} + 3 \times T_{vfy}) \times N_p = 545\,\mu\text{s}.$$

δV_{bl} is 1.3 V. The programming is entirely too slow because the number of program pulses N_p increases due to the limitation of V_{ch}, as seen in Fig. 17.

C. Proposed Method

The first page program is performed just like a two-level cell. ΔV_{th} of the "1" state is reduced down to 0.55 V by decreasing ΔV_{pp} to 0.25 V, as described above. As a result, the number of program pulses N_{p1} and the program time per page, T_{p1} are expressed as follows:

$$N_{p1} = 1 + (\Delta V_{th0} + \delta V_{pp})/\Delta V_{pp} = 11$$
$$T_{p1} = T_{load} + (T_{pulse} + T_{vfy}) \times N_{p1} = 234.5\,\mu\text{s}.$$

In the case of the "1" program the bitline is grounded, so there is no bitline voltage fluctuation, that is, δV_{ch} is 0 V. The verify read time T_{vfy} decreases to 4.5 μs due to the reduced bitline capacitance. Compared with a conventional four-level cell, the

Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614

program time decreases because there is no V_{ch} variation, and only one verify sequence is required.

At the second page program, the channel voltage of the selected memory cell is 0 V for the "3" program and 1.425 V for the "2" program, while the control gate is applied to V_{pp}. Fig. 16(c) shows the simulated write characteristics of the slowest cell. The "3" and "2" programs are performed at the same speed. As a result, the number of program pulses N_{p2} decreases. To make the maximum V_{th} the same as the conventional method, ΔV_{th} of the "2" and the "3" state is increased to 0.625 V, as shown in Fig. 7(b). Thus, ΔV_{pp} can be increased to 0.325 V, which further decreases N_{p2} and drastically improves program performance

$$N_{p2} = 1 + (\Delta V_{th0} + \delta V_{pp} + \delta V_{ch})/\Delta V_{pp} = 9.$$

The verify sequences decrease to two sequences ("2" and "3" verify) compared with the three sequences in the conventional method. This also decreases the program time per page T_{p2}

$$T_{p2} = T_{load} + (T_{pulse} + 2 \times T_{vfy}) \times N_{p2} = 236\,\mu\text{s}.$$

According to the preferential page select method described in Section VI, the first page has already been programmed before the second page program. Therefore, the program time of the second page is T_{p2}, not $T_{p1} + T_{p2}$. Consequently, whichever page is selected, any page can be programmed within 236 μs. Due to the reduced program pulses and verify sequences, a high programming speed of 236 μs/512 bytes or 2.2 Mbytes/s can be obtained, which is as much as 2.3 times faster than the conventional method, as shown in Fig. 17.

IX. Read Access Time

In NAND flash memories, the cell read current is as small as 1 μA. Therefore, the read access time is determined by the bitline capacitance and the delay time of wordline boosting has little effect on the access time. In the proposed cell array, the read operation is also accelerated because the bitline capacitance is halved. The read access time is 13.5 μs. This is 40% shorter than the conventional four-level cell.

X. Conclusion

A multipage cell architecture is proposed, allowing for both the precise control of the threshold voltage of a memory cell and fast programming without area penalty. A high programming speed of 236 μs/512 bytes or 2.2 Mbytes/s can be obtained, which is 2.3 times faster than the conventional method. A small die size can be achieved with a newly developed compact four-level column latch circuit. The preferential page select method is also proposed to improve the data retention characteristics. The IC error rate decreases by as much as 43%, and a highly reliable operation can be realized. The proposed architecture is a promising candidate for future highly reliable, high-speed programming and high-density multilevel NAND flash memories.

Acknowledgment

The authors would like to thank Dr. J. Miyamoto, K. Ohuchi, Dr. K. Sakui, Dr. R. Shirota, and K. Imamiya for their continuous encouragement. They also would like to express their gratitude to Dr. G. J. Hemink for his useful discussions and support.

References

[1] M. Ohkawa *et al.*, "A 98 mm^2 3.3V 64Mb flash memory with FN-NOR type four-level cell," in *ISSCC Dig. Tech. Papers,* Feb. 1996, pp. 36–37.
[2] K. D. Suh *et al.*, "A 3.3V 32Mb NAND flash memory with incremental step pulse programming scheme," in *ISSCC Dig. Tech. Papers,* Feb. 1995, pp. 128–129.
[3] K. Takeuchi *et al.*, "A multi-page cell architecture for high-speed programming multi-level NAND flash memories," in *Symp. VLSI Circuits Dig. Tech. Papers,* June 1997, pp. 67–68.
[4] T. Tanaka *et al.*, "A 3.4-Mbyte/s programming 3-level NAND flash memory saving 40% die size per bit," in *Symp. VLSI Circuits Dig. Tech. Papers,* June 1997, pp. 65–66.
[5] T. S. Jung *et al.*, "A 3.3V 128Mb multi-level NAND flash memory for mass storage applications," in *ISSCC Dig. Tech. Papers,* 1996, pp. 32–33.
[6] M. Bauer *et al.*, "A multi-level-cell 32 Mb flash memory," in *ISSCC Dig. Tech. Papers,* Feb. 1995, pp. 132–133.
[7] K. Imamiya *et al.*, "A 35 ns-cycle time 3.3V-only 32Mb NAND flash EEPROM," in *ISSCC Dig. Tech. Papers,* Feb. 1995, pp. 130–131.
[8] J. K. Kim *et al.*, "A 120 mm^2 64Mb NAND flash memory achieving 180ns/byte effective program speed," in *Symp. VLSI Circuits Dig. Tech. Papers,* June 1996, pp. 168–169.
[9] G. J. Hemink *et al.*, "Fast and accurate programming method for multi-level NAND flash EEPROM's," in *Symp. VLSI Technol. Dig. Tech. Papers,* June 1995, pp. 129–130.
[10] K. Takeuchi *et al.*, "A double-level-V_{th} select gate array architecture for multi-level NAND flash memories," in *Symp. VLSI Circuits Dig. Tech. Papers,* June 1995, pp. 69–70.
[11] ——, "A double-level-V_{th} select gate array architecture for multi-level NAND flash memories," *IEEE J. Solid-State Circuits,* vol. 31, pp. 602–609, Apr. 1996.
[12] H. Niijima, "Design of a solid-state file using flash EEPROM," *IBM J. Res. Develop.,* vol. 39, no. 5, pp. 531–545, 1995.
[13] T. Tanaka *et al.*, "A quick intelligent programming architecture and a shielded bitline sensing method for 3V-only NAND flash memory," *IEEE J. Solid-State Circuits,* vol. 29, pp. 1366–1373, Nov. 1994.

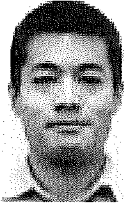
Ken Takeuchi was born in Yokohama, Japan, on November 12, 1967. He received the B.E. and M.E. degrees in applied physics from the University of Tokyo, Japan, in 1991 and 1993, respectively.

In 1993, he joined the Toshiba ULSI Research Laboratories, Toshiba Corporation, Kawasaki, Japan. In 1996, he transferred to the Toshiba Microelectronics Engineering Laboratory, Toshiba Corporation, Yokohama, Japan. Since 1993, he has been working on the circuit design of high-density flash memories.

Tomoharu Tanaka was born in Yamaguchi, Japan, on July 8, 1962. He received the B.E. and M.S. degrees in applied physics from the University of Tsukuba, Japan, in 1985 and 1987, respectively.

He joined the Toshiba Research and Development Center, Kanagawa, Japan, in 1987. Since then, he has been working on the circuit design of high-density EEPROM's. In 1996, he transferred to the Toshiba Microelectronics Engineering Laboratory, Toshiba Corporation. In 1997, he transferred to the Semiconductor System Engineering Center, Toshiba Corporation, Ofuna, Japan.

Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614



Toru Tanzawa was born in Yamanashi, Japan, on February 4, 1968. He received the B.E. degree in physics from Saitama University, Japan, in 1990 and the M.S. degree in physics from Tohoku University, Japan, in 1992.

In 1992, he joined the Toshiba Research and Development Center, Kanagawa, Japan. Since then, he has been working on the circuit design of high-density flash memories. In 1996, he transferred to the Toshiba Microelectronics Engineering Laboratory, Toshiba Corporation, Yokohama, Japan.

Kingston Technology Company, Inc., et al. EX1069
Kingston Technology Company, Inc., et al. v. Vervain, LLC IPR2025-00614
11 of 11