2.1 A 3.3V 1Gb Multi-Level NAND Flash Memory with Non-Uniform Threshold Voltage Distribution

Taehee Cho, Young-Taek Lee, Euncheol Kim, Jinwook Lee, Sunmi Choi, Seungjae Lee, Dong-Hwan Kim, Wook-Kee Han, Young-Ho Lim, Jae-Duk Lee, Jung-Dal Choi, Kang-Deog Suh

Samsung Electronics, Kyunggi, Korea

Although multi-level program cell NAND flash memory (MLC) is a promising mass storage device due to its low bit cost, it has been deprived of its market by single-level program cell NAND flash memory (SLC). This results from the difficulties of design and the low program speed of MLC. A 1Gb MLC with 2b per cell focuses on optimization of the multi-level cell V_{th} distributions, the key of MLC design for high program speed.

The 2b NAND flash memory having a 3-step programming sequence requires 3 times longer program time than that of a 1b NAND flash memory. To accelerate program speed, a 4x program mode (4xPGM) is adopted by dividing the cell array into 4 banks, giving 1.6MB/s program speed. Figure 2.1.1 is a die micrograph showing chip architecture. Each bank has 512B unit program depth compatible with the unit of conventional SLC and is organized in 8k columns by 16k rows. The banks are arranged in 2x2 matrix with page buffers at one side (center) to minimize die size and skew of control signals. The input control buffers are near input control pads, which are far from internal noise sources such as page buffers and high-voltage pumps.

The 4xPGM needs simultaneous charging and discharging of the 32k bitlines. Simulations show that peak current, where the bitline swing level is V_{cc}, is ~0.5A without rising and falling control of bitlines. Figure 2.1.2 shows the page buffer. The V_{refp} (V_{refn}) of REF_p (REF_n) transistor raises (lowers) the bitline level to V_{cc} (0V) in 2µs. In the beginning of bitline-setup, the 32k bitlines are charged up to V_{cc} through the REF_p with that slope. After 2µs, the bitlines of the cells to be programmed "1" are supported by the "1"-latched nodes and the "0"-program bitlines are discharged by the "0"-latched nodes. Consequently, the peak current is <60mA. In addition, time division sense-and-latch setup is used. Program and verify timing is similar to that shown in Reference 1 except that the S_{opt} is 0V when the chip operates as 512Mb SLC.

V_{th} of the programmed cell is influenced by adjacent wordline interference as the word line pitch is scaled down. Figure 2.1.3a is a schematic diagram showing wordline coupling with 0.5pF capacitance. During program operation, the string select line (SSL) of a selected block is set to V_{cc} and, inhibited bitlines, are set to V_{cc}, so SSL transistors are shut off. The channel underneath the selected cell is localized once V_{pass} is applied. When the V_{pgm} is applied to selected wordline (W/L_{15}), the SSL is coupled to W/L_{15} and is boosted to V_{cc} + 1.4V if no effort to reduce the V_{pgm} slope is made. Although the string transistor remains sub-threshold, the localized channel quickly loses its charge because the channel capacitance is more than 100 times smaller than the bitline capacitance. Therefore, inhibited cells are severely disturbed by V_{pgm}. A voltage-ramping circuit solves the problem. Figure 2.1.3b shows the change of channel potential underneath a selected cell to be inhibited with ramping. The ramping circuit has 5µs slope when V_{pgm} is applied to select wordline. As a result, the SSL coupling is reduced to 0.4V to have lower than a few pA sub-threshold current of the string select transistor, which drastically reduces the cell V_{th} shift due to wordline coupling.

The random order of page programming significantly shifts the cell V_{th} [1]. Another important effect is the interplay of maximum cell V_{th} and read voltage. The higher the read voltage, the smaller the background pattern dependency becomes. In this process, however, retention failure limits the maximum allowable read voltage to 5.5V. Figure 2.1.4 shows that the maximum V_{th} should be <2.6V to maintain the shift of cell V_{th} <1V for 5.5V read voltage. Cell V_{th} distribution must be as wide as possible because narrow cell V_{th} distribution degrades program performance and increases data retention failures. 2.6V maximum cell V_{th} is suitable.

Load current sensing is no longer practical because the small cell size reduces the worst-on-cell current to only 0.5µA, which is insufficient to drive the bitline within given sensing time. Instead, bitline precharge-and-sensing is adopted [2]. The scheme sequentially controlls the SBL level by 1.6V, 0V, and 1.0V (Figure 2.1.2). During sense, discharge of the precharged bitlines causes undershoot of the pocket-p-well (p-pwell). This positive body biases cells, which may mislead the sense-and-latches into recognizing a slightly underprogrammed cell as programmed. This is studied using the substrate model shown in Figure 2.1.5a. Figure 2.1.5b shows the noise decay time with the number of p-pwell straps. With only one p-pwell strap, the peak noise level is about 0.4V and decays to 0.05V within ~4µs, shifting the cell V_{th} by about 0.1V. There is a trade-off between chip size and the reduction of p-pwell noise. 15 p-pwell straps are included in this chip to reduce the p-pwell noise at the expense of 2% chip size.

In addition to the noise mentioned above, program disturbance, floating gate disturbance, and charge loss must be considered. It is inevitable that V_{th} distribution widths of the third ("01") and the highest ("00") states are widened because the erased state ("11") undergoes high start program voltage without experiencing the second ("10") state. Charges are induced on the floating gate when an adjacent cell is being programmed, which is measured to shift up the cell V_{th} by 0.1V. Charge loss by read cycling lowers the cell V_{th}. For the highest state, this shift amounts to 0.3V.

Target cell non-uniform threshold voltage distributions (NUTVD) of the 1Gb MLC are shown in Figure 2.1.6. The margins of neighboring states are so small that the NUTVD widths of each state must be kept narrow. Incremental step pulse programming with 0.15V step pulse is used for 0.3V NUTVD [3]. This chip also operates as a 512Mb SLC with 4.8MB/s program rate by cutting a fuse. The chip features are summarized in Figure 2.1.7.

References:
[1] Jung, T. S. et al., "A 3.3V 128Mb Multi-Level NAND Flash Memory for Mass Storage Applications," ISSCC Digest of Technical Papers, pp. 32-33, Feb. 1996.
[2] Tanaka, T. et al., "A Quick Intelligent Page-Programming Architecture and a Shielded Bitline Sensing Method for 3V-Only NAND Flash Memory," IEEE Journal of Solid-State Circuits Vol. 29, No. 11, pp. 1366-1373, Nov. 1994.
[3] Suh, K. D., et al., "A 3.3V 32Mb NAND Flash Memory with Incremental Step Pulse Programming Scheme," IEEE Journal of Solid-State Circuits Vol. 30, No. 11, pp. 1149-1156, Nov. 1995.

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 26,2024 at 02:02:36 UTC from IEEE Xplore. Restrictions apply.

Figure 2.1.1: Die micrograph.

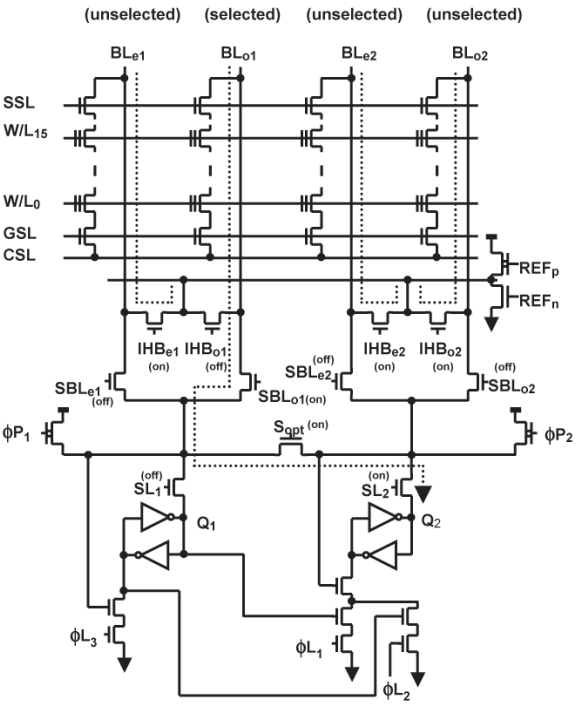

Figure 2.1.2: The sense-and-latch pagebuffer unit with single-bit-per-cell option transistor.

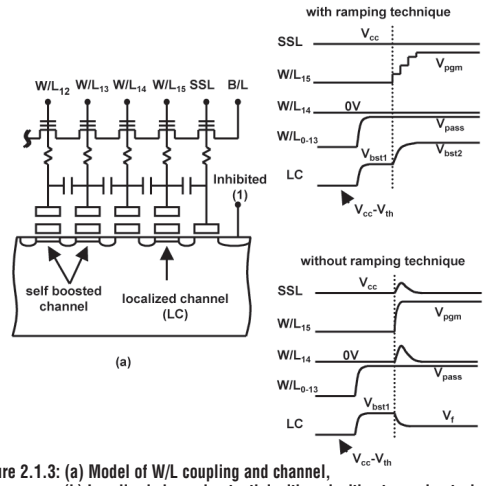

Figure 2.1.3: (a) Model of W/L coupling and channel,
(b) Localized channel potential with and without ramping technique.

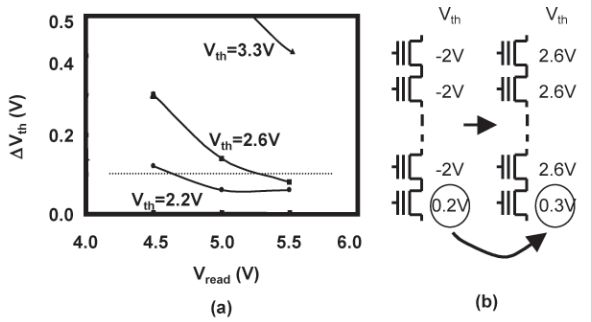

Figure 2.1.4: (a) Measured background pattern dependency,
(b) Cell V_{th} shift.

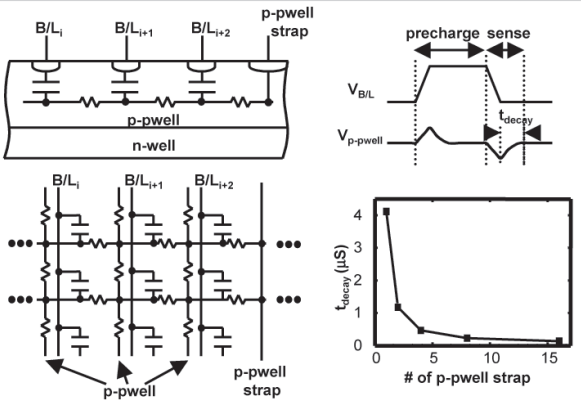

Figure 2.1.5: (a) Pocket pwell (p-pwell) noise model,
(b) Simulation result.

Continued on Page 424

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 26,2024 at 02:02:36 UTC from IEEE Xplore. Restrictions apply.

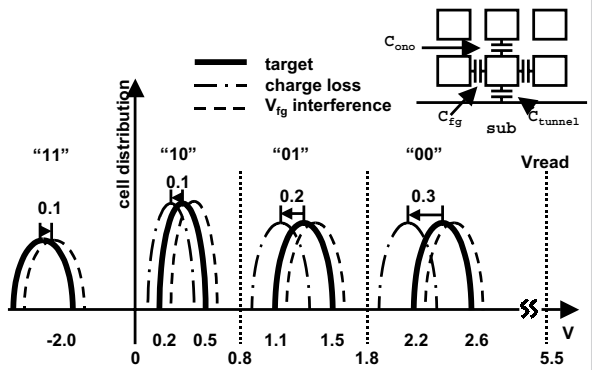

Figure 2.1.6: Threshold voltage distribution for 4 states.

Process	0.15µm CMOS, STI, triple-well, 2-metal
Tunnel oxide	8nm
Interpoly dielectric	15nm (effective)
Gate oxide	35nm (H.V), 8nm (L.V)
Cell size	0.14µm^2 (effective)
Chip size	116.7mm^2
Read access time	20µs transfer/50ns burst cycle
Block erase time	2ms
Page program time	1.2ms (1.6MB/s @ 2k mode)

Figure 2.1.7: 1 Gb NAND flash parameters and characteristics.

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 26,2024 at 02:02:36 UTC from IEEE Xplore. Restrictions apply.