Program Cache Busy Time Control Method for Reducing Peak Current Consumption of NAND Flash Memory in SSD Applications

Se-Chun Park, You-Sung Kim, Ho-Youb Cho, Sung-Dae Choi, Mi-Sun Yoon, Tae-Yun Kim, Kun-Woo Park, Jongsun Park, and Soo-Won Kim

In current NAND flash design, one of the most challenging issues is reducing peak current consumption (peak ICC), as it leads to peak power drop, which can cause malfunctions in NAND flash memory. This paper presents an efficient approach for reducing the peak ICC of the cache program in NAND flash memory — namely, a program Cache Busy Time (tPCBSY) control method. The proposed tPCBSY control method is based on the interesting observation that the array program current (ICC2) is mainly decided by the bit-line bias condition. In the proposed approach, when peak ICC2 becomes larger than a threshold value, which is determined by a cache loop number, cache data cannot be loaded to the cache buffer (CB). On the other hand, when peak ICC2 is smaller than the threshold level, cache data can be loaded to the CB. As a result, the peak ICC of the cache program is reduced by 32% at the least significant bit page and by 15% at the most significant bit page. In addition, the program throughput reaches 20 MB/s in multiplane cache program operation, without restrictions caused by a drop in peak power due to cache program operations in a solid-state drive.

Keywords: NAND flash memory, solid-state drive, peak current consumption, bit line, cache program.

I. Introduction

With the aggressive scaling down of the minimum feature size of memory bit cells, the capacity of NAND flash memory is drastically increasing, expediting NAND flash memory bit growth. Despite this merit, bit-line (BL) capacitances, which are shared by memory bit cells, are increasing abruptly [1] since the height of BLs is not reduced to maintain low resistances.

To support high-speed programming operations in NAND flash memories, BLs should be pre-charged in a very short time. However, since the target biasing of the write operation is much higher than that of the read operation, the peak current consumption (peak ICC) during the programming operation is becoming one of the largest concerns in low-power NAND flash memory design. Many previous research ideas have focused on suppressing peak ICC [1]–[6]; however, one of the difficulties encountered with these approaches is that the cache program operation is not seriously considered. Furthermore, cache I/O burst write current (ICC4W) is still a large problem in high-speed NAND flash memory design. Particularly, in the case of multiple concurrently operated NAND flash memories (for example, high-speed mass data storage modules such as solid-state drives (SSDs)), peak ICC is multiplied by the number of NAND flash memories that are operated concurrently. Figure 1 shows a typical SSD hardware architecture, which contains a large number of NAND flash memories and an SSD controller. In Fig. 1, the SSD controller can operate eight channels NAND chips concurrently with four-way interleaving to improve system performance. To improve the performance of the SSD, it is important to increase programming performance since programming operations are slower than reading operations in a NAND flash memory. However, due to the higher BL pre-charge target biasing of the

Manuscript received Nov. 5, 2013; revised Apr. 30, 2014; accepted May 17, 2014.

Se-Chun Park (sechun.park@sk.com), You-Sung Kim (yousung.kim@sk.com), Ho-Youb Cho (hoyoub.cho@sk.com), Sung-Dae Choi (sungdae.choi@sk.com), Mi-Sun Yoon (misun.yoon@sk.com), Tae-Yun Kim (taeyun3.kim@sk.com), and Kun-Woo Park (kunwoo.park@sk.com) are with the Flash design group, SK hynix, Seoul, Rep. of Korea.

Jongsun Park (jongsun@korea.ac.kr) and Soo-Won Kim (corresponding author, ksw@asic.korea.ac.kr) are with the School of Electrical Engineering, Korea University, Seoul, Rep. of Korea.

 © 2014 ETRI
http://dx.doi.org/10.4218/etrij.14.0213.0537

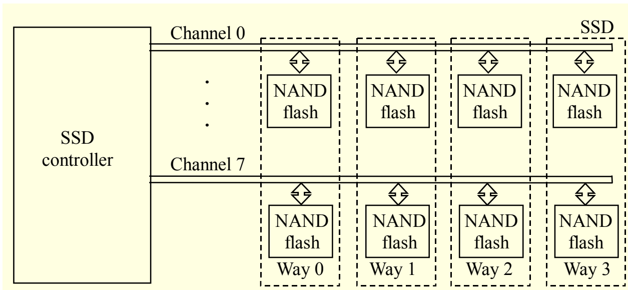


Fig. 1. Typical hardware architecture of an SSD.

write operation, an increase in program performance gives rise to a significantly larger peak ICC. The write performance of an SSD with interleaving is expressed as [1]

$$Performance_SSD = N \times Performance_NAND, \quad (1)$$

where N is the number of NAND flash memories operated concurrently (that is, the number of channels) and *Performance_NAND* is the write speed of a single NAND flash memory chip. In the SSD operation, since peak ICC increases as N increases, the maximum N is restricted by an ICC constraint [1]. One of the most well-known strategies for improving *Performance_NAND* is to use a cache program operation [7]. Here, since ICC increases with ICC4W, a large N gives rise to a large peak ICC [7]. The proposed program Cache Busy Time (*t*PCBSY) control method can effectively reduce the peak ICC of a cache program without restricting N [7].

II. Related Technologies

In NAND flash memories, various techniques for reducing peak ICC have been proposed [1]–[6]. In [1], a selective BL pre-charge, source-line program, and an intelligent interleaving scheme are proposed. In this work, a selective BL pre-charge scheme eliminates unnecessary BL pre-charging and the intelligent interleaving scheme avoids peak ICC through a power detector in the multi-wave interleaving operation. In [2], the drivability control of a BL pre-charge driver by reference voltage and bias slope are addressed. In [3], a two-step BL pre-charge technique (that is, in the first step, all BLs are pre-charged, and the BLs of the selected page are pre-charged in the second step) is introduced. In [4], the sequential sensing concept is addressed, which enables a BL to pre-charge only once in a multilevel sensing. An adaptive code selection scheme and a smart pre-charge algorithm are also introduced in [5] and [6], respectively. Although peak ICC is reduced during the program operation stage, peak ICC reduction with a cache program operation is not considered in [5]–[6].

III. Conventional Cache Program Method

Figure 2(a) shows the concepts of the conventional cache program [7]. In the beginning of the cache program, the data from the first page are loaded to the cache buffer (CB), which is referred to as the "Load" operation in the figure. Next, the data from the first page are transferred from the CB to the main buffer (MB) (labelled as "Transfer" in the figure). Then, the data from the first page are programmed to memory cells using the MB ("Program") operation. At the same time, the data from the second page are loaded to the CB ("Load" operation). The equation for the cache program performance of a NAND flash memory is as follows:

$$Performance_NAND = (Page\ size) / (t\text{PROG} + t\text{LOAD}). \quad (2)$$

As an example, in the case where a program time per page (*t*PROG) is 500 µs and the page size is 16 KB with I/O speed of 166 MB/s in NV-DDR mode, simple arithmetic indicates that the data load time (*t*LOAD) is 100 µs. Equation (2) shows that *Performance_NAND* is 27.3 MB/s. When a cache program operation [7] is employed, *Performance_NAND* can be improved by up to 32.8 MB/s, since *t*LOAD is hidden by *t*PROG. In (2), the "Transfer" time is ignored because it is approximately 1 µs. Generally, the ICC of the cache program is composed of array program current (ICC2) and ICC4W. In the example shown in Fig. 2, since the "Program" and "Load" operations are performed simultaneously, peak ICC is increasing. The problem is further aggravated when a high-speed I/O scheme is employed with a NAND flash memory, since this causes ICC4W to increase. This is one of the

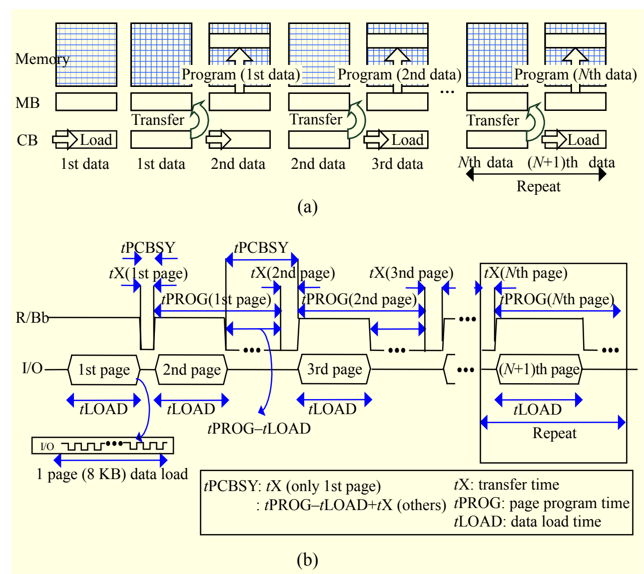


Fig. 2. Conventional cache program: (a) description of cache program and (b) timing diagram of cache program.

obstacles encountered when circuit designers try to design a high-performance NAND flash memory. Figure 2(b) shows a timing diagram that is relevant to Fig. 2(a). In the figure, the low state of R/Bb represents the busy status of the NAND flash memory. In a cache program operation, the Open NAND Flash Interface specification defines the period of time where a NAND is in "low state" mode as a *t*PCBSY period, during which cache data cannot be loaded to a CB.

IV. Proposed *t*PCBSY Control Approach

The main idea of *t*PCBSY control is that data from the second page can be loaded to the CB ("Load" operation in Fig. 2) when ICC2 becomes smaller through controlling *t*PCBSY. Since programming the NAND flash memories exploits Fowler–Nordheim (FN) tunnelling [8] and self-boost program inhibit schemes [9], ICC2 depends on the bias condition of the BLs. The ICC2 characteristic can be efficiently exploited in our approach to reduce the peak ICC. Figure 3 exhibits an even/odd BL structure. In the program operation, the BLs of the unselected page (BLo_1, BLo_2, and BLo_3 in Fig. 3) and the BLs of the completed program cells of the selected page (BLe_2 in Fig. 3) are pre-charged to the VDD (on-die power supply level) to inhibit the programming of the cells. Then the BLs of the incomplete program cells of the selected page (BLe_1 and BLe_3 in Fig. 3) were pre-charged to 0 V to program the cells using FN tunnelling [8]. In the NAND flash memory in Fig. 3, for equipotential BLs (BLo_2, BLe_2, and BLo_3 in Fig. 3), the effects of $C_{BL\text{-}BL}$ are ignored because it has the same potential at the two terminals (electrode) of. capacitance ($C_{BL\text{-}BL}$). Therefore, ICC2 can be minimized in the last program pulse because most of the cells are programmed

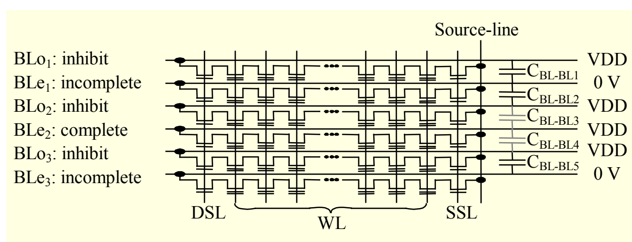


Fig. 3. NAND flash memory cell.

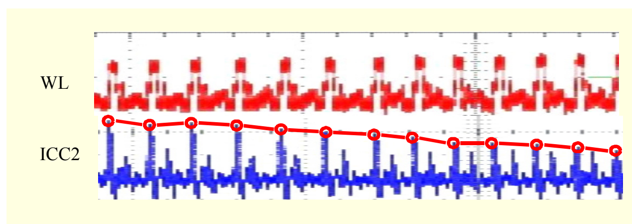


Fig. 4. Characteristics of ICC2.

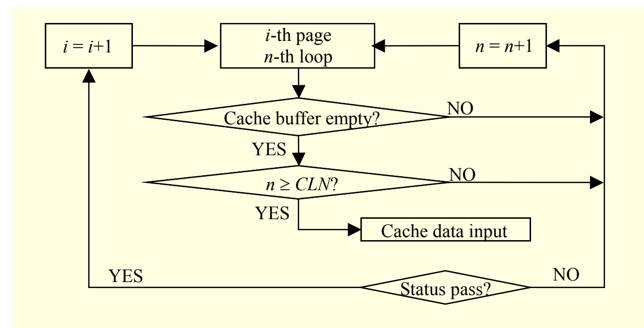


Fig. 5. Algorithm of proposed scheme.

Likewise, ICC2 is maximized in the first program pulse since most of the cells are not yet programmed. Figure 3 shows an example case, where $C_{BL\text{-}BL3,\,4}$ are ignored and $C_{BL\text{-}BL1,\,2,\,5}$ are effective. Figure 4 shows the measurements of ICC2 that were taken during programming of the most significant bit (MSB) page. Figure 5 shows the algorithm of the *t*PCBSY control method. In the cache program phase, after the *n*th program-pulse loop of the *i*th page is finished, the micro controller (MC) determines whether the CB is empty. If the CB is empty, the MC compares the loop number of the program pulse with the cache loop number (CLN). The CLN is a variable. It represents the program-pulse loop number of the cache program. The CLN indicates when data is loaded into the CB during the cache program phase. When *n* is greater than the CLN, MC sets *t*PCBSY to allow data insertion in the CB. The CLN is stored in an internal resistor, and the MC refers to the CLN to load the cache data. Here, the CLN is determined as the largest possible number, since the ICC2 peak is minimized in the last program pulse; however, in the case where the "Load" operation of the (*i*+1)th page isn't completed until finishing the "Program" operation of the *i*th page, the cache program performance degrades because *t*LOAD cannot be hidden by *t*PROG, as shown in (2). Nevertheless, a decrease in the performance of the cache program because *t*LOAD cannot be hidden by *t*PROG does not matter, for *t*LOAD has been decreasing with the evolution of high-speed I/O schemes in NAND flash memory. Particularly in a NAND flash memory, the number of program pulses is decreased due to the program/erase endurance cycle. In the strictest sense (that is, in the case where *t*LOAD is severely decreased), the endurance margin will limit the CLN. However, the endurance margin is small enough to operate the proposed scheme without performance degradation. Figure 6 shows a timing diagram showing a comparison of the proposed and conventional schemes' cache programs. In both schemes, cache data are loaded to the CBs when the CBs are empty ("Load" operation); however, it is only in the proposed scheme that peak ICC2 is decreased.

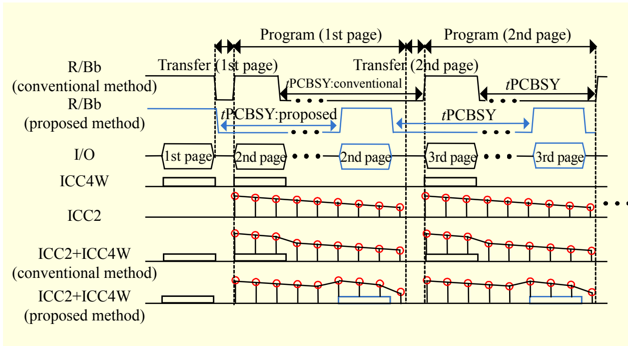
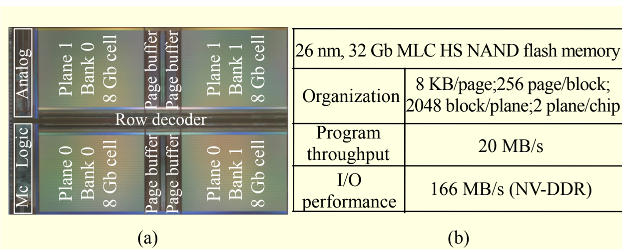


Fig. 6. Timing diagram showing a comparison of the proposed and conventional schemes' cache programs.

V. Measurement Results

Figure 7 shows a microphotograph and key features of the 26 nm 32 Gb high-speed (HS) MLC NAND flash memory. To evaluate the *t*PCBSY control method for reducing peak ICC in the cache program, we measured ICC during the operation of the cache program with CLN values of three and five. Figure 8 shows the measured ICC values in the cache programming of the least significant bit (LSB) page. Figure 9 shows the summary of the peak ICC improvements. As shown in the figure, the proposed *t*PCBSY control method achieves a peak ICC reduction of 32% on the LSB page and a peak ICC reduction of 15% on the MSB page.



26 nm, 32 Gb MLC HS NAND flash memory	
Organization	8 KB/page;256 page/block; 2048 block/plane;2 plane/chip
Program throughput	20 MB/s
I/O performance	166 MB/s (NV-DDR)

Fig. 7. (a) Microphotograph and (b) key features of flash memory device.

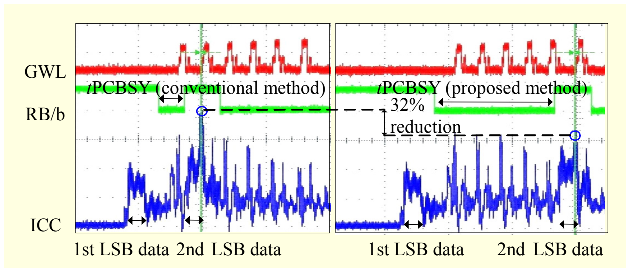


Fig. 8. Plot of experimental results.

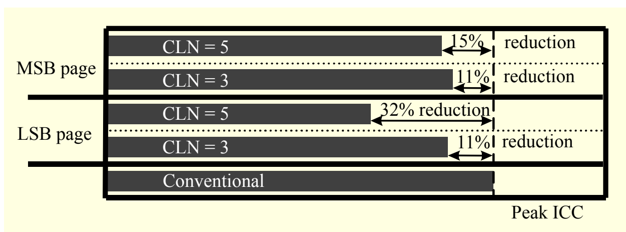


Fig. 9. Summary of experimental results.

VI. Conclusion

Since the peak ICC is multiplied by the number of channels, the *t*PCBSY control method for reducing the peak ICC in the cache program operation is essential for high-speed interface applications with multi-channel organization, such as in SSD architecture. In this paper, we proposed an efficient approach for reducing the peak ICC of the cache program in NAND flash memory — namely, the *t*PCBSY control method. It enables the SSD controller to operate a multiplane cache program, without malfunctions caused by a drop in peak power due to multiple concurrent cache program operations in SSD applications.

References

[1] K. Takeuchi, "Novel Co-Design of NAND Flash Memory and NAND Flash Controller Circuits for Sub-30 nm Low-Power High-Speed Solid-Stage Drives (SSD)," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, Apr. 2009, pp. 1227–1234.

[2] T. Cho et al., "A 3.3 V 1 Gb Multi-level NAND Flash Memory with Non-Uniform Threshold Voltage Distribution," *Proc. IEEE ISSCC*, San Francisco, CA, USA, Feb. 7, 2001, pp. 28–29.

[3] T. Cho et al., "A Dual-Mode NAND Flash Memory: 1-Gb Multilevel and High-Performance 512-Mb Single-Level Modes," *IEEE J. Solid-State Circuits*, vol. 36, no. 11, Nov. 2001, pp. 1700–1706.

[4] C. Trinh et al., "A 5.6 MB/s 64 Gb 4 b/Cell NAND Flash Memory in 43 nm CMOS," *Proc. IEEE ISSCC*, San Franciso, CA, USA, Feb. 8–12, 2009, pp. 246–247.

[5] C. Lee et al., "A 32-Gb MLC NAND Flash Memory with Vth Endurance Enhancing Schemes in 32 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, Jan. 2011, pp. 97–106.

[6] K. Fukuda et al., "A 151-mm^2 64-Gb 2 Bit/Cell NAND Flash Memory in 24-nm CMOS Technology," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, Jan. 2012, pp. 75–84.

[7] K. Imamiya et al., "A 125-mm^2 1-Gb NAND Flash Memory with 10-MByte/s Program Speed," *IEEE J. Solid-State Circuits*, vol. 37, no. 11, Nov. 2002, pp. 1493–1501.

[8] R.H. Fowler and L. Nordheim, "Electron Emission in Intense Electric Fields," *Proc. Royal Soc.*, May 1, 1928, pp. 173–181.

[9] K. Suh et al., "A 3.3 V 32 Mb NAND Flash Memory with Incremental Step Pulse Programming Scheme," *Proc. IEEE ISSCC*, San Francisco, CA, USA, Feb. 15–17, 1995, pp. 128–129.