Virtium
Solid State Storage and Memory

Application Note

AN07-0819-01

Improving Endurance with Garbage Collection, TRIM and Wear Leveling

Introduction

Prior to the advent of solid state drives (SSD), hard disk drives (HDD) were the primary storage medium. HDDs use heads and platters to store data and can be written to and erased an almost unlimited number of times, assuming no failure of mechanical or circuit components (one of the main drawbacks of HDDs for industrial applications). SSDs, on the other hand, have no moving parts, thus are much more durable; however, they only support a finite number of program/erase (P/E)[1] operations. For single-level cell (SLC) NAND, each cell can be programmed/erased 60,000 to 100,000 times.

Wear-leveling and other flash management techniques, while will use some of the cycles, provide protection for the NAND and the data on the SSD. These are part of the Write amplification factor (WAF)[3]. The final number is typically translated as Terabytes Written (TBW)[2]. For (industrial-grade) multi-level cell (MLC) and triple-level cell (TLC), each cell can be programmed/erased 3,000 to 10,000 times. These conventional NAND flash types use what is referred to as “floating gate” technology whereby during writing operations, tunnel current passes through the oxide layer of the floating gate in the NAND cell (Figure 1.).

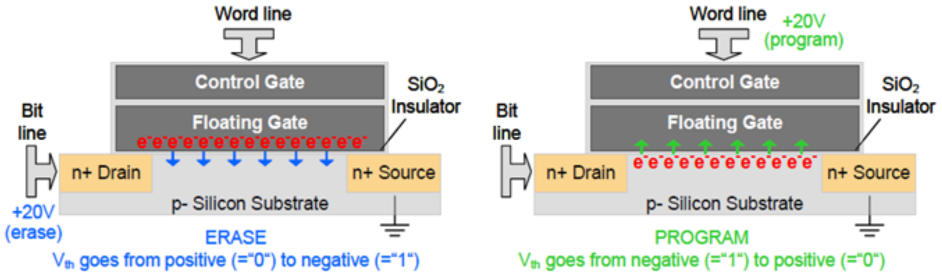


Figure 1. NAND Flash Floating Gate

This occurrence, known as electron tunneling causes the oxide layer to degrade with each charged/trapped electron; the more the cell is programmed (written to or erased), the faster it degrades, eventually causing the cell block to wear out where it then cannot be programmed any longer and turns into a read-only device. Typically, this wear-out is preceded by a decrease in performance as well as an increase in uncorrectable bit errors. Effective flash management techniques and SSD monitoring/reporting tools can dramatically improve SSD endurance and provide predictive maintenance insights to better manage SSDs in service.

Flash Management Solutions

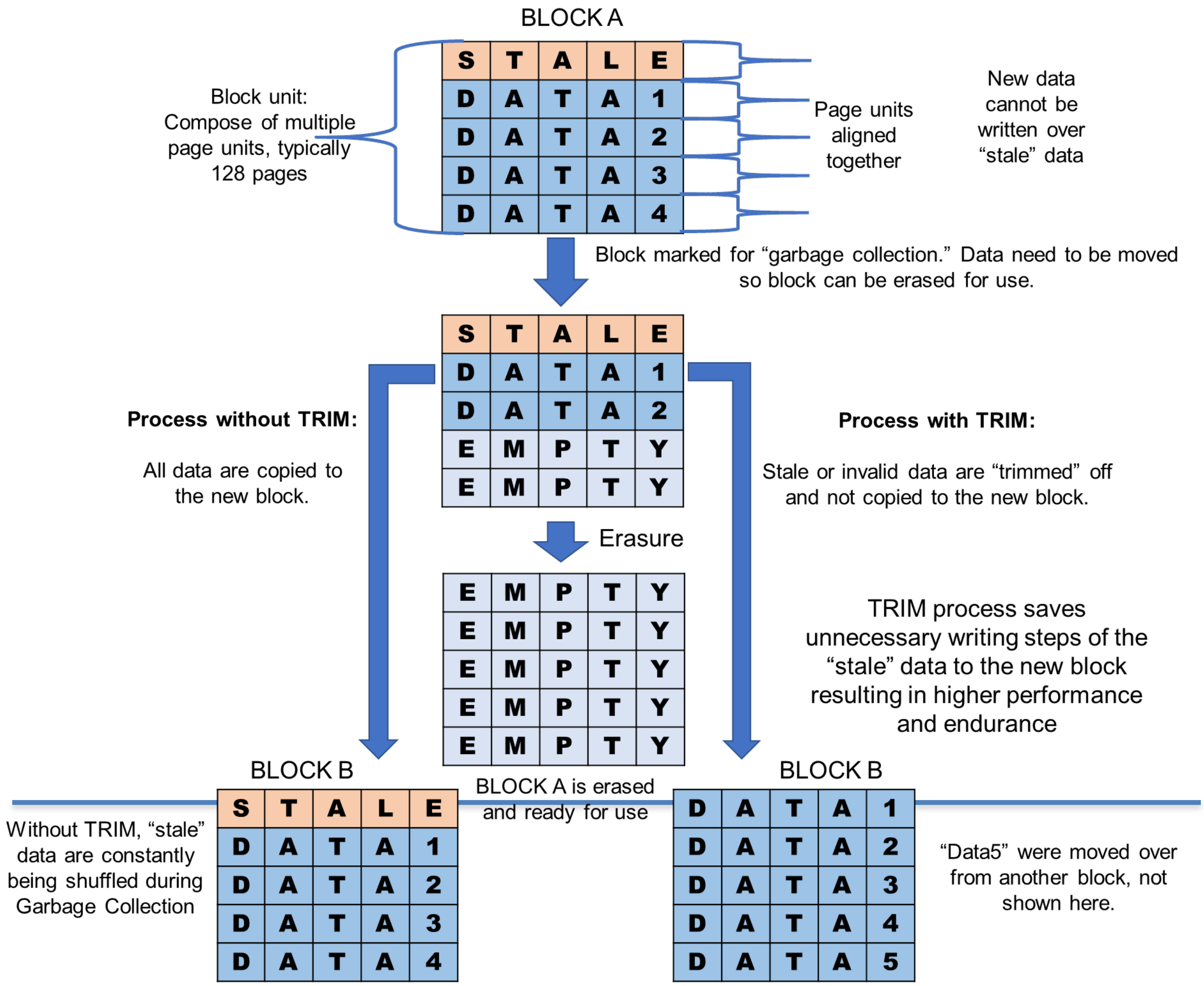
Garbage Collection

In order for the SSD to maintain its write/read performance, blocks in the NAND chip must be free of data so they can readily be available to accept new data to be written. When the SSD’s initial store of free pages runs out, the SSD needs to erase old blocks before writing new ones. While writing operations take place in "page" units that consist of multiple NAND strings, erasing can only be performed in "block" units which are made up of multiple pages. Therefore, any page that needs to be erased, the entire data (good and bad) within that block would have to be erased.

Garbage collection is a process of having the controller search through its inventory of written pages for pages that have been marked as "stale," and move them to a new cell block, (stale or “invalid data” are data that were written to but needed to be modified by the OS, and therefore needs to be rewritten). Garbage collection occurs during lull and does not interfere with the controller’s performance, thereby optimal writing speed during normal operations is maintained.

Trim

To reduce the number of unnecessary data writes that increase the WAF[3], an ATA command called TRIM is utilized to enable the operating system to notify the SSD and mark which data page is “stale” and tells the SSD to ignore those invalid data during the garbage collection process. This eliminates any unnecessary copying of stale data pages during the process which reduces the total number of program/erase cycles to the NAND and prolongs the life of the SSD. Overall, TRIM makes an SSD’s garbage collection more efficient. An SSD is no longer forced to save pages belonging to stale data files, resulting in more efficiently managed storage space available for data. Figure 2. shows an overview of garbage collection with and without TRIM implemented.



BLOCK A
S T A L E
D A T A 1
D A T A 2
D A T A 3
D A T A 4
Block unit: Compose of multiple page units, typically 128 pages
Page units aligned together
New data cannot be written over “stale” data
Block marked for “garbage collection.” Data need to be moved so block can be erased for use.
S T A L E
D A T A 1
D A T A 2
E M P T Y
E M P T Y
Process without TRIM:
All data are copied to the new block.
Process with TRIM:
Stale or invalid data are “trimmed” off and not copied to the new block.
Erasure
E M P T Y
E M P T Y
E M P T Y
E M P T Y
E M P T Y
TRIM process saves unnecessary writing steps of the “stale” data to the new block resulting in higher performance and endurance
BLOCK B
S T A L E
D A T A 1
D A T A 2
D A T A 3
D A T A 4
Without TRIM, “stale” data are constantly being shuffled during Garbage Collection
BLOCK A is erased and ready for use
BLOCK B
D A T A 1
D A T A 2
D A T A 3
D A T A 4
D A T A 5
“Data5” were moved over from another block, not shown here.

Figure 2. Simplified visualization of how data are moved and rewritten

Wear Leveling

Within the NAND flash, there are two types of data, static and dynamic. Static data contains information such as operating systems, executable and user files that is rarely updated. It may be read frequently, but seldom changes. Dynamic data, on the other hand, changes often and therefore requires frequent rewriting/erasing.

If data was to be reprogrammed/erased back and forth multiple times between the same cell blocks, uneven wear would occur resulting in premature damage to the targeted blocks. To prevent this untimely block damage, a technique called “wear leveling” is employed to ensure evenly distributed wear to all cells. This involves a diversification of write operations so that they are not concentrated on a specific block. Figure 3. shows an example of an uneven wear of the cell blocks due to lack of wear leveling. There are two types of wear leveling: static and dynamic.

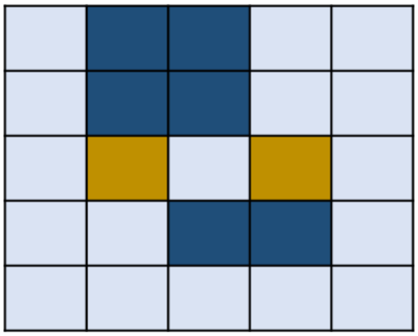


Figure 3. Uneven wear of blocks

Legend:

⬜Not used 🟦Somewhat used 🟨Damaged

If data were to move back and forth between the same group of blocks, premature damage would occur, rendering those blocks useless.

Dynamic Wear Leveling

Dynamic wear leveling pools only data-free or erased available blocks to be wear-leveled. The one with the lowest erase count will be targeted for new data writes. Consequently, the same group of blocks get wear leveled, which are typically those that were occupied by dynamic data.

Static Wear Leveling

Blocks containing static data typically don’t get rewritten often and therefore will have lower erase counts compared to blocks utilized by dynamic data. With static wear leveling implemented, data in any block with the lowest erase count will be transferred to a block with higher erase count. The block is then erased and considered for the next data write. Static wear leveling essentially tracks the cycle count of “ALL good blocks” and attempts to evenly wear out each block by selecting the available block with the least wear count for writing operations.

Wear leveling is a basic technology for SSD devices, but its effectiveness depends largely on the algorithm that is used for controlling and spreading the rewrite operations. Depending on the file size ratio of static to dynamic data, static wear leveling can substantially improve the life of the SSD compared to dynamic wear leveling alone. Virtium SSDs use NAND controllers that support both dynamic and static wear leveling that enables highly accurate and finely tuned control to achieve superior results.

Table 1. is taken from the SMART attribute of an actual test results from one of Virtium’s SSDs. The data from the SMART attribute table shows the controller effectively spreading the wear to all the NAND blocks with an average erase count of 4375, with the highest and lowest erase counts of 4545 and 4318 respectively.

164	Attribute -	8928414 Total Erase Count
165	Attribute -	4545 Maximum Erase Count
166	Attribute -	4318 Minimum Erase Count
167	Attribute -	4375 Average Erase Count
168	Attribute -	20,000 Max NAND Erase Count from specification

Table 1. SMART attribute showing results of effective wear leveling

Conclusion

SSD is a storage device that is necessary to optimize. NAND flash wears out due to the repeated effects of the program/erase cycle, therefore reducing the number of P/E counts can lengthen the endurance of the SSD. The use of the TRIM command helps to optimize the capacity of an SSD by allowing garbage collection process to ignore the stale data from repeated programming/erasing, resulting in faster data writes and longer SSD life. The life of the SSD can further significantly increase with the implementation of wear leveling technology. With wear leveling, practically all blocks are evenly worn out.

Definitions

1. **NAND Program/Erase (P/E) Cycles –** the number of program/erase cycles supported by NAND, based on a given error correction (ECC) capability of the SSD controller, at an assumed temperature (40°C, 104°F) and an assumed data retention requirement (1 year for SLC and MLC; 3 months for TLC).

2. **Terabytes Written (TBW) –** the total amount of data that can be written into an SSD over its lifetime. For example, if your SSD is rated for 100 TBW, it means you can write 100 TB into it before it reaches its warranted lifetime. TBW is defined by the following equation:

$$TBW = \frac{SSD\ capacity\ * flash\ cell\ life}{WAF}$$

SSD capacity is specified in GB and flash cell life is specified by the manufacturer

3. **Write Amplification Factor (WAF)** – the amount of data written to NAND versus amount of data written by the host. This value is greater than 1, assuming no data compression, and is the result of a mismatch in the page size of the NAND (the minimum write unit) and the block size of the NAND (the minimum erase unit). Write amplification is determined by firmware algorithms and workload. It can range from an ideal state of 1, where data comes across the interface in large file, sequential transfers and is aligned to NAND flash page boundaries; to a worst case of block size divided by page size. In general, the more random the write workload, the higher the write amplification. WAF is defined by the following equation:

$$WAF = \frac{Flash\ writes}{Host\ writes}$$

Virtium
Solid State Storage and Memory

30052 Tomas | Rancho Santa Margarita, CA 92688
Phone: 949-888-2444 | Fax: 949-888-2445

www.virtium.com