ApproxFTL: On the Performance and Lifetime Improvement of 3D NAND Flash based SSDs

Jinhua Cui, Youtao Zhang, *Member*, *IEEE*, Liang Shi, *Member*, *IEEE*, Chun Jason Xue, *Member*, *IEEE*, Weiguo Wu, Jun Yang

***Abstract*—Three-dimensional (3D) NAND flash is one of the most prospective advances in flash memory industry. While 3D flash improves cell density and reduces lithography cost through die stacking, it suffers from severe program disturbance, which leads to significant performance and lifetime degradation for 3D flash based SSDs.**

To address the above challenge, we propose ApproxFTL, an approximate-write aware flash translation layer design, that uses *approximate-write* operations to store error-resilient data of modern applications. By reducing the maximal threshold voltage and tightening the guard bands between multi-level cell states, approximate write operations not only finish early but also exhibit large disturbance reduction, which can be exploited to alleviate disturbance in physical blocks that save both precise and approximate data. ApproxFTL maximizes the disturbance mitigation through approximate-write aware data placement, wear leveling, and garbage collection enhancements. Our experimental results show that ApproxFTL, while preserving high data quality, improves the read and write response time of flash accesses by 41.38% and 45.64% on average, respectively, and extends the lifetime of 3D flash based SSDs by 5.75% when comparing to the state-of-the-art.

***Index Terms*—3D Flash memory, Approximate storage, Reliability.**

I. Introduction

THREE-DIMENSIONAL (3D) NAND flash is one of the most prospective advances in flash memory industry [1]. By stacking flash cells in vertical direction, 3D flash achieves significant density improvement and lithography cost reduction. However, a 3D flash page has more neighboring pages than a planar page does. As such, programming a 3D flash page disturbs more pages in the same block[1] while a programmed page is subject to more disturbing write operations, i.e., when its neighboring pages are programmed. With fast technology scaling, program disturbance is projected to be one of the major challenges in 3D flash [2].

Most schemes developed for mitigating program disturbance in 3D flash were proposed at the hardware level, e.g., those that redesign the cell structures [3] and those that redesign the ECC circuit [4]. At the system level, Wang *et al.* proposed to reorder writes sent to physical blocks to minimize inter-block disturbance [5]. Chang *et al.* proposed to reorder writes to pages within one physical block [6]. Chang *et al.* proposed to mitigate disturbance by redesigning the program operation and sequence [7]. Chang *et al.* proposed to split a large 3D block to sub-blocks with lower disturbance [8].

In this paper, we propose ApproxFTL, an approximate-write aware FTL (flash translation layer) to mitigate program disturbance in 3D flash and address the performance and lifetime degradation of 3D flash based SSDs. A recent work that is close to our design is to trade data accuracy for improved solid state memory access performance [9]. By exploiting the error resilience in modern applications, Sampson *et al.* proposed to reduce the number of write steps when writing error resilient data. The technique was designed for PCM but can be adapted to multi-level cell (MLC) flash. A major difference between their approach and ours is that the former cannot mitigate program disturbance, which is the main design goal of our work. The following summarizes our contributions.

- We propose to realize approximate flash write, i.e., page programming, by reducing the maximal threshold voltage and tightening the guard bands between MLC flash states. The approximate-write not only speeds up flash accesses to approximate data, but also greatly reduces the disturbance to neighboring pages.
- We then propose ApproxFTL, an approximate-write aware FTL design, to fully exploit the benefits of approximate-write operations. It integrates three enhancements in FTL.
  (i) We enhance the baseline data allocation to prioritize the checkerboard-style allocation of approximate and precise pages in physical blocks, which effectively minimizes the disturbance on precise pages in each block;
  (ii) We enhance the baseline wear leveling to exploit the

[1]In this paper, we focus on VG (vertical gate) 3D flash cell architecture that has one block spread across multiple layers. The design is applicable to other 3D cell architectures.

Manuscript received July 07, 2017; revised October 14, 2017; accepted November 26, 2017. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1000303 and Grant 2016YFB0201800, and by the Joint Research Fund for Overseas Chinese, Hong Kong and Macao Young Scientists of the National Natural Science Foundation of China under Grant 61628210, and in part by the National Natural Science Foundation of China under Grant 91630206 and Grant 61672423 and Grant 61772092, and in part by National Science Foundation of the United States under Grant CCF-1718080, and in part by National 863 Program under Grant 2015AA015304. The work of J. Cui was supported by the Chinese Scholarship Council under Grant 201606280098.

J. Cui and W. Wu are with the school of Electronic and Information Engineering, Xi'an Jiaotong University, Shaanxi, 710049, China (E-mails: cjhnicole@gmail.com; wgwu@xjtu.edu.cn).

Y. Zhang is with the Computer Science Department, University of Pittsburgh, Pittsburgh, PA 15260, USA (E-mail: zhangyt@cs.pitt.edu).

S. Liang is with the College of Computer Science, Chongqing University, Chongqing, China, and with the Key Laboratory of Cyber Physical Society Credible Service Computing, Ministry of Education, Chongqing, China (E-mail: shi.liang.hk@gmail.com).

C. Xue is with the Department of Computer Science, City University of Hong Kong, Hong Kong (E-mail: jasonxue@cityu.edu.hk).

J. Yang is with the Electrical and Computer Engineering Department, University of Pittsburgh, Pittsburgh, PA 15261, USA (E-mail: juy9@pitt.edu).

lifetime benefit from reducing maximal threshold voltage. By tracking the wearing effect based on the data allocation pattern of different blocks, we distribute approximate and precise writes proportionally to each physical block;
(iii) We enhance the baseline garbage collection such that, at the garbage collection time, it moves valid approximate and precise pages in two batches to appropriate physical blocks for maximized disturbance reduction.

- We evaluate the proposed schemes and compare them to the state-of-the-art. Our experimental results show that ApproxFTL, while preserving high data quality, improves the read and write response time of flash accesses by 41.38% and 45.64% on average, respectively, and extends the lifetime of 3D flash based SSDs by 5.75% when comparing to the state-of-the-art.

The rest of this paper is organized as follows. Section II presents the background and related work. Section III discusses the approximate page programming that mitigates program disturbance in 3D flash. Section IV elaborates the design details of ApproxFTL. Section V lists the experiment settings and analyzes the results. We conclude the paper in Section VI.

II. Background and Related Work

A. 3D NAND flash memory

Three-dimensional (3D) NAND flash is a type of flash that achieves density improvement through die stacking. A single etching can punch through multiple layers and connect the cells from these layers [1]. 3D NAND flash memory can be classified into two types based on the flowing direction of the string current — VC (vertical channel) and VG (vertical gate). In VC 3D flash, e.g., BiCS [10] and V-NAND [11], the current flows vertically while in VG 3D flash, e.g., 3DVG [4], [12], the current flows horizontally so that the gate signals are vertically shared. Studies have shown that VG flash has better pitch scalability due to its smaller minimal cell size and well-controlled interface [6], [13]. In this work, we elaborate our design using VG 3D flash while the proposed schemes are applicable to VC 3D flash.

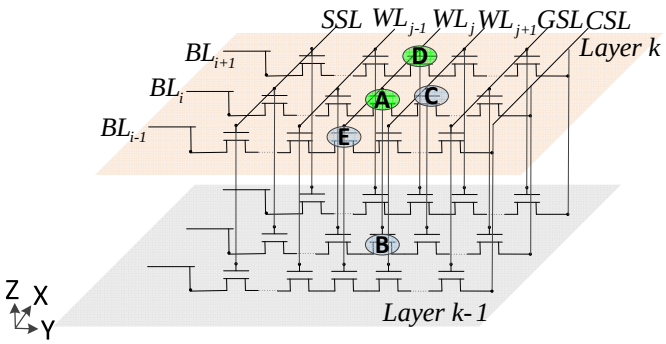


Fig. 1. The circuit schematic of 2-layer VG 3D NAND flash.

Fig. 1 presents a block equivalent circuit schematic of two-layer 3D flash. A physical block in 3D flash consists of $N \times L$ pages where L is the number of layers of the flash module and N is the number of pages in each layer that belong to the block. The pages from different layers are vertically stacked so that their corresponding wordlines (WLs) are spliced together along the Z direction. The layout within each layer is traditional, that is, a bitline (BL) has contacts with all wordlines (WLs) in one layer for the given block while the bits from one page are laid out along the WL direction.

Program disturbance to be mitigated in the paper. As the number of layers is increasing in 3D flash, the deteriorated program disturbance is a extremely critical design issue in 3D flash memory [6], [7]. That is, the threshold voltage of a programmed flash page may be unintentionally shifted when programming its neighboring pages, which is mainly attributed to parasitic capacitances coupling effect between adjacent ones [7]. In the example, when programming cells A and D, we may disturb their neighboring cells, e.g., cell B and C. Studies have shown that this type of disturbance becomes much larger in 3D flash than that in 2D flash. Yeh *et al.* showed that the disturbance along Y direction is about 1.33 times of that along Z direction in the same chip [14]. In this paper, we focus on mitigating capacitive coupling effect based program disturbance, which is the same as most of existing 3D disturbance mitigation designs [6], [15], [16].

By adapting the threshold voltage shift models in planar flash [17], [18], the threshold voltage shift on the cell being disturbed in 3D flash memory, due to the capacitive coupling effect based program disturbance, can be modeled as following.

Given a cell s at *(i, j, k)* where *i* is cell offset within one flash page (along BL-to-BL, or X direction), *j* is the page index on one layer (along WL-to-WL, or Y direction), *k* is the layer index (along Layer-to-Layer, or Z direction). We have $0 \le i <$ PSIZE, $0 \le j <$ N; $1 \le k \le$ L, and PSIZE, N, L are the the total number cells in one page, the number of pages in one layer, and the number of layers, respectively. For discussion clarity, we skip the cells in the top/bottom layers and at the outer lines of each block. They have fewer neighboring cells and are also handled in the experiments.

$$\begin{aligned} \Delta V_{(s)} &= \sum_{\forall t} \frac{C_{(s)(t)} \ \Delta V_{(t)}}{C_{total}} \\ s &= (i,j,k) \\ t &\in \{(i,j-1,k),(i,j+1,k), \\ &\quad (i,j,k-1),(i,j,k+1)\} \end{aligned} \tag{1}$$

where $C_{\texttt{total}}$ is the total capacitance of the victim cell; $C_{(s)(t)}$ is the coupling capacitance between the victim cell s and the its neighboring cell t. For example, cell *(i,j,k+1)* is the one on top of the victim cell. The four neighboring cells along Y and Z directions have non-negligible disturbance while the disturbance effects along WL direction and diagonal directions are often small and neglected [14]. $\Delta V_{(t)}$ indicates the threshold voltage change before and after programming cell t. Cell s and t are referred to victim (or disturbed) and disturbing cell, respectively, in the rest of the paper.

A victim cell s gets disturbed if any of its neighboring cells t along Y and Z directions is programmed after programming s, according to Equation (1). For a typical sequential pro-

grammed order that programs cells layer-by-layer and then WL-by-WL within each layer, a cell may be disturbed up to two times. For a random programming order, a cell may be disturbed up to four times. The disturbance effects accumulate and lead to errors if $\Delta V_{(t)}$ becomes too larger.

Given that capacitance between cells depends on their distance, this type of program disturbance is projected to increase dramatically with technology scaling. The disturbance was obviously observed for planar flash at 1x-nm technology node and below, which has become one of the major scaling challenges [19]. For 3D flash chips that currently adopt larger technology node (>40nm), the disturbance begins to manifest as being significant [6], [7]. While cell optimizations help to mitigate its severity temporarily [11], [20], with 3D integration fast approaching its stacking limits, future 3D flash chips have to scale to smaller technology node, which face severe disturbance, similar as that in 2D flash.

Other disturbance sources. There exist two more types of disturbance. One is the channel coupling based program disturbance. For example, when programming cells A and D in Figure 1, other cells on the same WL, e.g., cell E, refrered to as inhibited cells, may be disturbed. $V_{\texttt{prog}}$ and V_{cc} are applied to the WL and BL of each inhibited cell, respectively. The inhibited channel is boosted during programming to avoid unintentional threshold voltage shift. Unfortunately, one WL is physically spliced together across multiple layers along Z direction, thus 3D stacking leads to more disturbance errors on inhibited cells [2].

The other disturbance comes from flash read operations. For the unselected cells in the same block during read, their WLs and BLs are applied $V_{\texttt{pass}}$ and GND. While one read contributes little disturbance, a 3D flash block contains much more pages than its 2D counterpart such that there is large accumulated disturbance for long lived hot pages.

The significance of these two types of disturbance may amplify with technology scaling. We leave the design of effective schemes on mitigating them in our future work.

B. Approximate Computing

Approximate computing is an emerging computing paradigm that exploits the inherent error resilience in many modern applications, such as audio, video, and image processing applications [21]. For example, it is often not noticeable to have a small number of errors in rendering video streams; and having a small number of bad pixels in high resolution images usually does not affect image classification. In order to ensure the overall quality of service (QoS) for these applications, data is often divided into *critical data* and *non-critical data*, with the help from either the programmer [21] or a compiler [22]. While errors may appear in non-critical data, the critical data are kept precise.

Most of approximate computing proposals exploit error-resilience in cache and main memories and/or within approximation-aware microarchitecture. Liu *et al.* proposed to reduce the refresh frequency of DRAM banks that save non-critical data, which not only saves refresh energy but also improves performance [21]. Esmaeilzadeh *et al.* developed approximation-aware hardware to efficiently support approximate programming model [23].

Extending approximate computing to solid state memories enables the construction of approximate storage. Sampson *et al.* proposed to use fewer steps when programming MLC PCM (Phase Change Memory) [9], which can be extended to MLC flash to improve the performance of writing non-critical data. Guo *et al.* observed that the bits of encoded image have non-uniform error resilience and proposed to store them in separate regions with different precision guarantee [24]. Djordje *et al.* computed bit-level reliability requirements for encoded videos, and showed that reliability can be traded for density improvement in video storage [25].

A main difference between approximation storage schemes and ApproxFTL is that the former achieves performance and energy saving benefits by speeding up the processing of approximation data. ApproxFTL further targets at reducing program disturbance in physical blocks so that it benefits the processing of both precise and approximate data.

A critical problem in approximate computing is to evaluate the quality of the approximate output. Early designs adopted static evaluation and sampling schemes. The recent advances proposed online evaluation. For example, Khudia *et al.* proposed to dynamically monitor output quality and adjust computation accuracy accordingly [26]. This is orthogonal to our design. Currently ApproxFTL adopts pessimistic static evaluation, which can be improved by online error monitoring for better performance.

C. The Flash Translation Layer (FTL)

The Flash Translation Layer (FTL) is a critical component in NAND flash based SSDs. It has been extensively studied for improved system performance and extended lifetime of SSDs. Most FTL designs [27], [28] partition hot and cold data, which helps to minimize write amplification. By integrating data compression [29] and de-duplication [30], the SSD lifetime may be prolonged with reduced flash writes. Optimizing garbage collection policies in FTL [31] helps to improve the average response time for flash systems.

ApproxFTL differs from these FTL designs in that it leverages approximate write to mitigate program disturbance in 3D flash and improve the performance.

III. Approximate Write on 3D Flash

Approximate computing is an emerging computing paradigm that requires the coordination across software and hardware layers. In particular, previous approximate computing approaches differentiate precise data and approximate data, which may be annotated either by the programmers or using an approximate computing aware compiler, e.g., Enerj [9]. ApproxFTL makes the similar assumption in implementing the flash memory based approximate storage — the annotation is passed together with each write request to the host interface logic (HIL) of the SSD.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2017.2782765, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

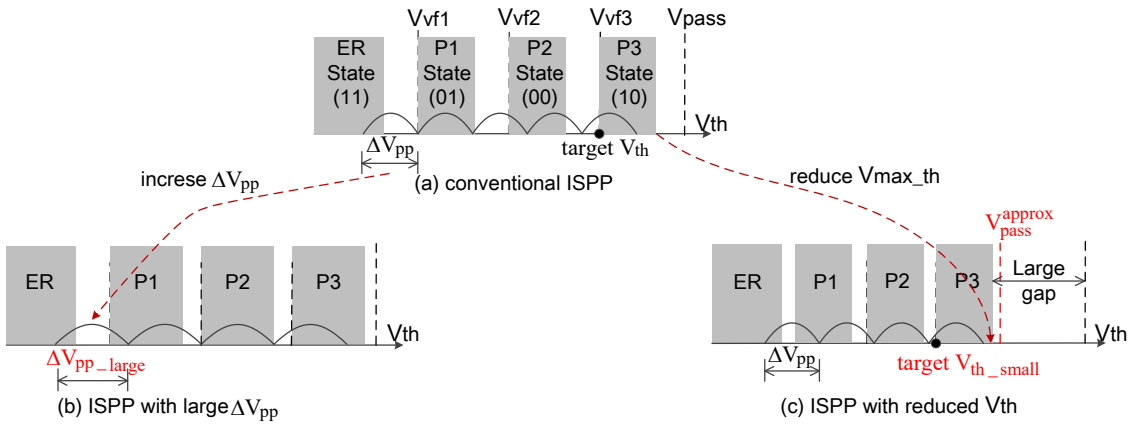


Fig. 2. The ISPP schemes for (a) the baseline (i.e., precise page programming); (b) approximate programming with large ΔV_{pp}; (c) approximate programming with reduced threshold voltage V_{max_th}.

A. ISPP based Flash Programming

Flash programming widely adopts the incremental step pulse programming (ISPP) strategy. A typical ISPP scheme consists of multiple steps with each step applying an increasingly biased programming voltage to raise the voltage of the cells being programmed. The programming voltage increment at each step, referred to as ΔV_{pp}, is often a constant. After each step, a verify operation is carried out to check if the cell voltage is above the target voltage level (target V_{th}) and determine if the programming process can be terminated. Fig. 2(a) illustrates the four target reference voltages of 2-bit MLC cells. The shaded range indicates the voltage window for each state while the white gap between two states indicates the guard band.

Studies have revealed that flash reliability and access performance depend on the adopted ISPP scheme. Given that cell voltage may fluctuate after programming due to disturbance and charge leakage, the voltages of a small number of flash cells may deviate into the ranges of their neighboring states, resulting in raw bit errors (RBEs) that demand ECC code to fix before returning the data to the user. On one hand, the more bit errors a flash page contains, the longer error correction process it takes during read, and the lower the read performance is. On the other hand, choosing a smaller ΔV_{pp} helps to achieve narrower voltage range for each state and wider guard bands, i.e., the data are more reliable. Due to larger guard bands, the same voltage shift leads to fewer errors so that the read is faster. However, it takes more steps and longer latency to finish the write operation.

B. Approximate Write for 3D Flash

Based on the interdependency between flash reliability and access performance, a simple implementation of approximate write is to use larger ΔV_{pp} than that in the baseline, which reduces the number of write steps and improves the write performance [9] [2].

[2]Sampson *et al.* proposed to reduce the number of write steps for PCM [9], here we adapt it to 3D flash for comparison.

As shown in Fig. 2(b), adopting a larger ΔV_{pp} shrinks the guard bands between MLC states, which leads to a significant increase of raw bit error rates (RBER). Our experiments show that the RBER changes from 10^{-8} when using conventional precise programming to 7.2×10^{-4} when using a large ΔV_{pp}, which is based on the split-page 3D VG NAND Flash chip [14] to carry out these Monte Carlo simulations (see Section III-C for more details). Given that such high RBER is beyond the correction capacity of an ECC code with reasonable space overhead, the ECC correction phase for a read operation may be skipped [32]. The quality of the output is ensured either by static analysis [22] or online error monitoring [26]. In summary, approximate write improves both read and write performance by trading the data accuracy.

In the following discussion, we use *precise/approximate write* to denote the ISPP that does-not/does relax data reliability, respectively. That is, we expect 10^{-8}/7.2×10^{-4} RBER, respectively. We use *precise/approximate page* to denote the data left in the physical page after precise/approximate write, respectively.

The simple approximate write, while improving the access performance to approximate data, has limited impact on accessing precise data. More importantly, it lacks the ability to mitigate the challenging program disturbance problem in 3D flash.

In this paper, we adopt an alternative approximate programming, which reduces the maximal threshold voltage V_{max_th}, and accordingly the reduced target voltage level V_{th_small} and the guard bands, keeping the same ΔV_{pp} as in the baseline, as shown in Fig. 2(c). Comparing to the scheme using large ΔV_{pp}, our V_{max_th}-reduction based approach has the following advantages.

- Reducing V_{max_th} helps to reduce program disturbance in physical blocks. According to Equation (1), the program disturbance depends on the cell voltage change before and after programming. Reducing V_{max_th} reduces the gratitude of the voltage change, indicating mitigated disturbance on the victim cells.
  Our study shows that we may reduce V_{max_th} by up to 38% while maintaining comparable RBER on approximate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2017.2782765, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

data as that using larger ΔV_{pp}, i.e., the RBER increases from 10^{-8} when using conventional precise programming to 7.2×10^{-4} in our approximate programming (see Section III-C for more details). Programming an approximate page with reduced V_{max_th} reduces its disturbance to neighboring cells proportionally, i.e., by up to 38% reduction.

- Reducing V_{max_th} helps to improve the write performance when writing either precise or approximate pages. Approximate write becomes faster because we reduce the guard bands so that it takes fewer steps to finish the ISPP programming.
  Whether we can speed up a precise write operation depends on how we are to program its neighboring pages. Let us assume pages are sequentially programmed in the physical block, and we use (*, j, k) to denote a flash page in layer k with page index j.
  (1) In the case if we may potentially write precise pages to (*, j+1, k) and (*, j, k+1), we have to be conservative in programming page (*, j, k), that is, it has the same performance as that in the baseline;
  (2) In another case, we may have decided only to allocate approximate pages to (*, j+1, k) and (*, j, k+1). Given that programming (*, j+1, k) and (*, j, k+1) exhibits 38% less disturbance on (*, j, k), we may slightly relax the reliability in writing precise page (*, j, k). We observe around 24% performance improvement.
  (3) In yet another case, we may decide to write those neighboring pages first and write (*, j, k) the last. For page (*, j, k), there is no future disturbing operation, while in traditional sequentially program case both (*, j+1, k) and (*, j, k+1) exhibit disturbance on page (*, j, k). We may more aggressively relax the reliability in writing precise page (*, j, k) than that of case (2). We may achieve maximized write relaxation, which is about 33% faster than the baseline.
  Given that the performance improvement of precise write depends on how the pages are allocated in the physical block, we are motivated to develop approximate-write aware data allocation. We will elaborate it in details in the next section.
- Reducing V_{max_th} helps to achieve less stress for the cells from an approximate page. Studies have shown that the wearing effect of flash memory depends on the erase voltage at erase time [33], [34]. When there is a reduction of V_{max_th}, we can reduce erase voltage proportionally such that the P/E cycle can be improved.

Reducing the maximal threshold voltage method has also attracted research attention by other previous works, which are used for traditional precise storage with improved performance/lifetime [34], [35]. But they need to lower the V_{max_th} voltage conservatively, where data are accommodated by the flash error correction capability. We would like to reduce the V_{max_th} voltage more aggressively in approximate write scenario, where its RBER may exceed the error correction capability. Moreover, using approximate-write operations to store error-resilient data of modern applications, we propose ApproxFTL, an approximate-write aware flash translation layer design to explore above advantages with different page allocation patterns. We will elaborate it in details in the Section IV.

A drawback of reducing V_{max_th} is that it reduces the retention reliability. In our design, only approximate pages may be affected, i.e., suffer from an increment of retention errors. Given approximate pages save non-critical data, our experimental study shows its impact is negligible. For precise pages, we use the same V_{max_th} as the baseline to avoid the retention issue for critical data.

It should be noted that approximate computing blurs the boundary for excluding traditionally defective writes. The RBERs of writing precise and approximate computing are 10^{-8} and 7.2×10^{-4}, respectively, exhibiting more than $1000\times$ difference. The RBER of the approximately written cells, when considering the process variations (PVs) and suffering from the worst-case write disturbance and retention charge loss would go beyond LDPC correction capability, which fails application execution under the traditional computing paradigm. However, the errors can generate acceptable application outputs under the approximate computing paradigm.

C. Modeling Details

In this section, we present the details of reliability model under different maximal threshold voltages. The following formulas are extended from [36], but different from [36], the 3D flash reliability model exploits program disturbance coming from Z direction.

At first, threshold voltage distribution of erase $P_e(x)$ and programmed states $P_p^{(k)}(x)$ can be expressed as:

$$P_e(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

$$P_p^{(k)}(x) = \begin{cases} \frac{1}{\Delta V_{pp}}, & if \quad V_p^{(k)} \le x \le V_p^{(k)} + \Delta V_{pp} \\ 0, & other \end{cases} \tag{3}$$

Program disturbance increases the threshold voltage of the victim cell systematically, where its threshold voltage shift is modeled in Equation (1). Its threshold voltage distribution follows:

$$P_{pd}(x) = \begin{cases} \frac{\alpha}{\sigma_d\sqrt{2\pi}} e^{-\frac{(x-\mu_d)^2}{2\sigma_d^2}}, & if\ |x-\mu_d| \le w \\ 0, & other \end{cases} \tag{4}$$

where the BL-to-BL and diagonal coupling ratio are often small and neglected, as explained before, and WL-to-WL and Layer-to-Layer coupling ratio is γ_y and γ_z, respectively.

The probability density function of threshold voltage with P/E-induced fluctuation, and retention time-induced fluctuation can be expressed as below.

$$P_{pe}(x) = \frac{1}{2\lambda} e^{-\frac{|x|}{\lambda}} \tag{5}$$

$$P_{rt}(x) = N(\mu_r, \sigma_r^2) \tag{6}$$

Then the final threshold voltage distribution of the k-th programmed state is obtained as:

$$P^{(k)}(x) = P_p^{(k)}(x) \otimes P_{pd}(x) \otimes P_{pe}(x) \otimes P_{rt}(x) \tag{7}$$

Therefore, RBER model can be computed as:

$$RBER = \sum_{k=0}^{q} \left(\int_{-\infty}^{V_p^{(k)}} p^{(k)}(x)dx + \int_{V_p^{(k+1)}}^{+\infty} p^{(k)}(x)dx \right) \tag{8}$$

UBER is typically required to be under 10^{-15} [39], and the relationship between UBER and RBER is:

$$UBER = \frac{\sum_{n=t+1}^{N_{cw}} \binom{n}{k} \cdot RBER^n \cdot (1 - RBER)^{(N_{cw}-n)}}{N_{user}} \tag{9}$$

The variables and notations in the reliably model are leveraged from [36], [37]. According to the JEDEC standard JESD47I.01 [38], Floating Gate(FG)/Charge Trap (CT) flash memory needs 1000 hour retention time in high-temperature retention bake, and the 10K maximal P/E cycling for MLC flash memory. On the basis of 3D VG NAND flash memory [14], the program disturbance contributes 0.2 and 0.15 shift through Y and Z direction, respectively, whereas we set ΔV_{pp} and $V_{\texttt{max\_th}}$ as 0.2, and 8, in the conventional ISPP scheme (Figure 2(a)). And In case of programming with reduced threshold voltage $V_{\texttt{max\_th}}$ (Figure 2(c)), we aggressively set ΔV_{pp} and $V_{\texttt{max\_th}}$ as 0.2, and 5, respectively.

IV. ApproxFTL Design

In this section, we present an overview of ApproxFTL and elaborate the three enhancements that we integrate at the FTL level.

A. An Overview

ApproxFTL is an approximate-write aware FTL that exploits $V_{\texttt{max\_th}}$-reduction based approximate page programming for maximized disturbance reduction and performance improvement in 3D-flash based SSDs.

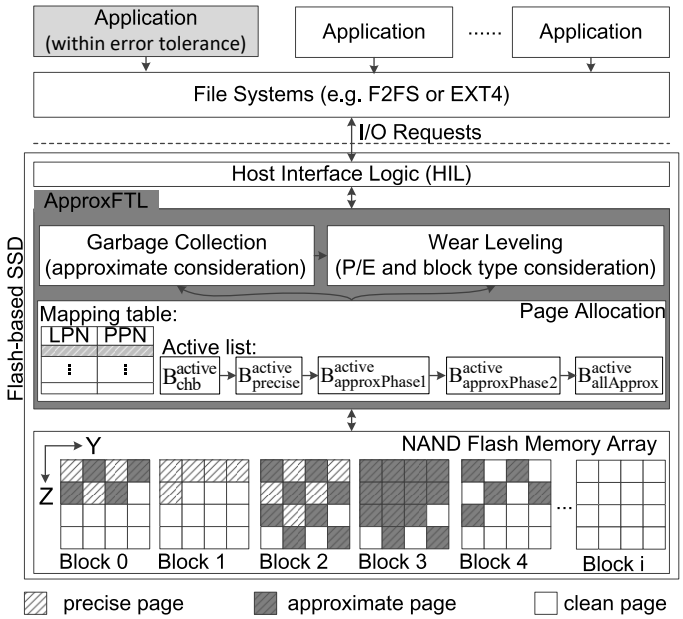


Fig. 3. An overview of ApproxFTL.

An overview of ApproxFTL is shown in Fig. 3. At the high level, an error-resilient application has its data partitioned into critical region and non-critical region, with the assistance from either manual tagging or compilation analysis [22]. The file system is enhanced to pass the approximate tag to the FTL in the SSD. That is, an I/O request sent to flash FTL is a quadruple

`(LPN, Size, Read/Write, ApproxFlag),`

where `LPN` is the starting logical page number of the data block, `Size` is the length of the data block, `ApproxFlag` denotes if the user data can be written using approximate write. FTL then strips the data block to multiple chips for achieving chip-level parallelism.

At the low level, an approximate-write controller is integrated inside the flash module. It receives the requests together with approximate tag from the enhanced FTL, and conducts the corresponding approximate and precise writes with enhanced ISPP programming.

At the middle level, we enhance the baseline FTL to exploit the disturbance mitigation inside each physical block. We first enhance the baseline page allocator, the component in FTL for assigning physical page number (PPN) to each incoming page write with its LPN. We develop a pattern-guided page allocator that determines the allocation pattern before allocation. By taking advantage of the pattern information, ApproxFTL effectively minimizes disturbance and speeds up both approximate and precise page writes.

We then enhance the baseline wear leveling algorithm with approximate-write awareness. Since page wearing effect depends on its maximal threshold voltage, we differentiate approximate and precise, track wearing effect based on block usage, and strive to proportionally distribute both types of write requests to each page.

We also enhance the baseline garbage collection to mitigate program disturbance. At the garbage collection time, ApproxFTL moves valid approximate and precise pages in two batches to active blocks for maximized disturbance reduction. In addition, we promote approximate pages to precise ones to prevent introducing unbounded precision errors.

Next, we elaborate the three major enhancements in ApproxFTL.

B. Pattern-guided Page Allocation

In Section 3, we have presented that the programming of precise pages can be speeded up. This is enabled when their neighboring locations are programmed with approximate pages. This motivates our pattern guided allocation strategy — when fetching a clean block for LPN-to-PPN mapping, we pre-determine its page allocation pattern, i.e., how approximate and precise pages are laid out in the block, and then strictly follow the pattern in the future mapping.

Page allocation pattern. In this paper, we adopt the following three page allocation patterns while more patterns will be exploited for further optimization.

- Checkerboard pattern. When there are about the same number of precise and approximate page writes, we allocate them interleavingly in both Y and Z directions, similar to

a checkerboard, e.g., block 0 in Fig. 3. We next discuss the disturbance reduction in the block by considering two different write orders.

The first write order is sequential interleaving — starting from the first page, we take approximate and precise writes *alternatively* and map them to physical locations in sequential order. By adopting this order, a precise page, e.g., page s that was just programmed, may be disturbed at most by two future approximate writes, one in Y direction and one in Z direction. In this case, the precise write can be speeded up, as discussed in Section 3. For an approximate page, it may be disturbed by up to two precise writes, which is the same as the baseline. The disturbance on approximate pages has little influence.

The second write order is two-phase interleaving — we first write all approximate pages and then write the precise pages to the block. During the phase of writing approximate pages, we skip locations that are allocated to precise pages. As a result, by the time when we write a precise page, all its neighboring pages (approximate pages) have been programmed. Since there is no future program disturbance to the precise page, we can speed up precise write by a larger amount comparing to that using sequential interleaving write order.

- All precise pattern. If there are significantly more precise page writes than approximate ones, or precise page writes arrive in a burst, we need to fetch physical blocks for precise writes only. We program pages in the block in sequential order such that each programmed page may be disturbed by two precise writes (at later times). This is the same as the baseline.

  For this pattern, we take the sequential order by pessimistically assuming all previously programmed pages are alive when we program the last page of the block. Other writing orders [6] may achieve better disturbance mitigation and performance improvement.
- All approximate pattern. If there are significantly more approximate page writes, we allocate physical blocks and map only approximate writes to them. Similar as above, we choose the sequential order when writing pages to the block.

LPN-to-PPN Mapping. We next elaborate the enhancement for the mapping from an incoming write (with LPN) to a PPN. Setting up the mapping is often simple in existing FTLs: the FTL keeps one active physical block, maps incoming writes (with different LPNs) to consecutive PPNs in the block, and fetches a new block after depleting the current one. With the recent advance in flash industry, e.g., the multi-stream technology in Samsung 840 Pro SSD [39], [40], the FTL maintains more than one active blocks such that a write can be mapped to a better choice.

Since our pattern-guided page allocation restricts the capability of mapping an incoming write to an active block, we propose to maintain at most five active blocks with different allocation patterns such that an incoming write can always be directed to the best block. The small number of active blocks has little impact on performance as shown in [39].

As shown in Fig. 4, we keep the following active blocks. A block is always tagged as being clean after erase, i.e., $B^{\texttt{clean}}$.

- The $B^{\texttt{active}}_{\texttt{chb}}$ block. This block is to take the checkerboard page allocation pattern, and the sequential interleaving write order. That is, it services approximate and precise writes alternatively. For example, if the last LPN page mapped to the block is an approximate write, the FTL has to map a precise LPN page before mapping another approximate page.

  After programming the last page of $B^{\texttt{active}}_{\texttt{chb}}$, the FTL converts it to $B^{\texttt{used}}_{\texttt{chb}}$ and allocates a clean block to $B^{\texttt{active}}_{\texttt{chb}}$.
- The $B^{\texttt{active}}_{\texttt{precise}}$ block. This block is to take the all precise pattern and sequential write order. This is the same as the baseline approach. The $B^{\texttt{active}}_{\texttt{precise}}$ block is converted to $B^{\texttt{used}}_{\texttt{precise}}$ after its last write, and then the FTL allocates a clean block to $B^{\texttt{active}}_{\texttt{precise}}$.
- The $B^{\texttt{active}}_{\texttt{approxPhase1}}$ block. This block is to take the checkerboard pattern and two-phase interleaving write order. The block is currently in the first phase, i.e., servicing approximate writes. Precise writes cannot be sent to this block.
- The $B^{\texttt{active}}_{\texttt{approxPhase2}}$ and/or $B^{\texttt{active}}_{\texttt{allApprox}}$ block. When $B^{\texttt{active}}_{\texttt{approxPhase1}}$ finishes the first phase, the FTL converts it to either $B^{\texttt{active}}_{\texttt{approxPhase2}}$ or $B^{\texttt{active}}_{\texttt{allApprox}}$. The former indicates that the block is now servicing precise writes while the latter indicates that the block continues to service approximate writes.

  After the above conversion, the FTL moves a clean block to $B^{\texttt{active}}_{\texttt{approxPhase1}}$. Similar as above, the $B^{\texttt{active}}_{\texttt{allApprox}}$ block is tagged as `used` after its last write. Note that $B^{\texttt{active}}_{\texttt{approxPhase2}}$ will be tagged as $B^{\texttt{used}}_{\texttt{chb}}$ after its last write, because it takes the checkerboard pattern as well.

At any given time, we have up to five active blocks — we always have one of the first three types and may be missing one or both of the last two active blocks types. Fig. 4 summarizes the life cycle of a physical block in ApproxFTL.

Algorithm 1: Active Block Selection

Input: Write requests.
Output: $B_{\texttt{current}}$, the selected active block.

```
repeat
    if request.ApproxFlag==B^active_chb.want then
        B_current=B^active_chb ;
    else if request.ApproxFlag==precise then
        if B^active_approxPhase2 exist then
            B_current=B^active_approxPhase2 ;
        else
            B_current=B^active_precise ;
        end
    else
        if B^active_approxPhase2 no exist then
            B_current=B^active_approxPhase1 ;
        else
            B_current=B^active_allApprox ;
        end
    end
until no write request;
```

Choosing the best block. Given an approximate or precise write may be serviced by more than one blocks, Algorithm 1

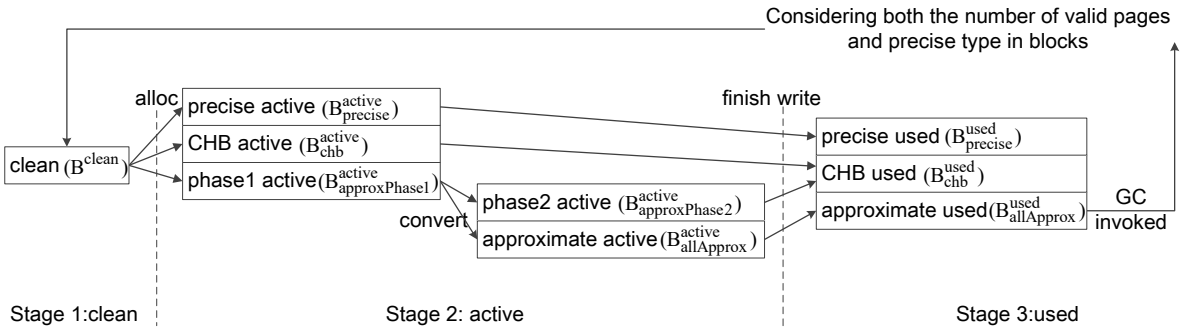


Fig. 4. The life cycle of flash memory blocks.

elaborates the details on how to choose the best one. For the writes from the file systems, we first check if it is the precision type that B^{active}_{chb} wants and gets serviced if it is a match. Otherwise, we direct approximate writes to $B^{active}_{approxPhase1}$ and $B^{active}_{allApprox}$, and precise writes to $B^{active}_{approxPhase2}$ and $B^{active}_{precise}$.

When we have all five active blocks, we pause mapping approximate writes to $B^{active}_{approxPhase1}$ until one of $B^{active}_{approxPhase2}$ or $B^{active}_{allApprox}$ depletes.

When we miss $B^{active}_{approxPhase2}$ and have all other active blocks, we prioritize the mapping of approximate pages to $B^{active}_{approxPhase1}$ (rather than $B^{active}_{allApprox}$). The goal is to proceed to the end of the first phase early such that $B^{active}_{approxPhase1}$ can be converted to $B^{active}_{approxPhase2}$ to service precise writes.

When we only have the first three block types. $B^{active}_{approxPhase1}$ is by default converted to $B^{active}_{approxPhase2}$ when it reaches the end of the first phase.

Separating Hot and Cold Data. The benefits of separating hot and cold data have been evaluated in several previous studies [41]–[50]. In this paper, this separation strategy is also implemented to improve the GC performance, and we categorize the hot/cold requests according to their read or write data sizes, i.e., small-sized approximate/precise requests are treated as hot ones. The size-based strategy was widely adopted to classify hot and cold data in recent studies [41]–[44]. Many such small-sized requests process metadata and thus are critical to system performance. While recent designs confirmed the effectiveness of this metric, data hotness can be defined differently [45]–[50]. The proposed designs in this paper are independent of the hotness definition and can be adapted to use other hotness metrics. All of the flash blocks are separated into two allocation pools, one hot pool for hot data and another cold pool for cold data. The hot data will be navigated to blocks in the hot pool, and vice versa.

C. Wearing leveling

In recent studies, Jeong *et al.* [33] and Shi *et al.* [34] revealed that reducing the maximal threshold voltage V_{max_th} enables the adoption of a lower erase voltage, which reduces the *effective wearing*. For example, in [34], when reducing V_{max_th} from 4.25V to 3.40V, erasing the block 100 times with lower voltage corresponds to 80 erases with the default voltage, i.e., the *effective wearing* is 0.8.

In this paper, we extend *effective wearing* to take advantage of the wearing benefits from approximate writes. In particular, for the blocks that only save approximate data, all pages share the same reduced V_{max_th} so that a lower erase voltage can be employed, which exhibits proportional *effective wearing* reduction, i.e., 0.62 of the baseline. For other blocks, we use the default erase voltage as at least some pages in these blocks have no V_{max_th} reduction.

ApproxFTL tracks the *effective wearing* of all physical blocks. Given a block b_i, ApproxFTL updates its *effective wearing* W_{b_i} at erase time using Equation (2). We only exploit the blocks that save approximate data in this paper.

$$W_{b_i} \quad += \quad \begin{cases} 0.62 & : \ if\ b_i\ is\ a\ B^{used}_{allApprox}\ block \\ 1.00 & : \ otherwise \end{cases} \tag{10}$$

While W helps to track the effective wearing at the block level, cells within a physical block have different wearing effects if the block adopt checkerboard page allocation pattern. Instead of developing expensive schemes to track at cell level, ApproxFTL takes a simple strategy that flips the type of page that can be allocated to its first physical page. That is, the first physical page, if it is mapped to service an approximate write this time, would service a precise write next time. The flag is checked, used and flipped only when the block is assigned to use checkerboard pattern. Using the block with other patterns leads to uniform stress of all cells.

At runtime, ApproxFTL tracks the effective wearing W of all physical blocks based on their programming pattern, and ideally returns the clean block with the smallest W if a new clean block is requested. We found it is often not necessary to track the smallest W, in the experiments, we identify the smallest W periodically and, when a clean block is requested, conduct a greedy search and return the first block whose W is not two times bigger. Our results show that this simple heuristic performs equally well while greatly reducing the sorting overhead.

D. Garbage Collection and Refresh

ApproxFTL converts active blocks to three types of *used* blocks after programming their last pages (Fig. 4). When there is a need to invoke garbage collection, a traditional collector would reclaim the block with the smallest number of valid pages, regardless of their page types, i.e., if these pages are approximate or precise pages. Such a garbage collector often leads to suboptimal results.

We enhance the baseline garbage collection as follows.

- *Victim block selection policy.* At the time to identify a victim block, we not only consider the total number of valid pages but also their approximate type. For blocks with the same number of valid pages, we prioritize the selection of blocks that have fewer valid precise pages. This is because, when moving the same number of valid pages, the smaller the number of precise pages is, the smaller the overall disturbance the move introduces to the system.
  In particular, if there is no active block $B^{active}_{approxPhase2}$, we demote the selection of $B^{used}_{precise}$ blocks as the latter contains all precise blocks. Moving valid pages from the block would copy them to $B^{active}_{precise}$, which has large overall disturbance.
- *Valid page copying policy.* When copying valid pages from the victim block, we first move all valid approximate pages and then the precise pages. Copying approximate pages in batch helps to create $B^{active}_{approxPhase2}$ if it does not exist. Valid precise pages can then be copied to this block rather than $B^{active}_{precise}$, which introduces minimized program disturbance.
- *Approximate page upgrade policy.* Since writing approximate pages tends to introduce uncorrectable errors in the page, copying these pages to new locations, if keeping using approximate write, may accumulate significant large amount of errors that fail the application. For this reason, we attach a flag to each block to record the largest number of approximate writes to any given page. We promote all approximate pages of the victim block to precise page if this flag reaches the threshold.
  For example, we set the flag of a block B_1 to 1 if the approximate pages of this block are written with data from the file system. Garbage collecting this block copies its approximate pages to a new block B_2, we set the B_2's flag to 2 if it is not bigger. Assuming our promotion threshold is 2, garbage collecting B_2 at a later time shall promote all approximate pages to precise ones, which helps to prevent introducing unbounded errors to the approximate pages.

An approximate page accumulates retention errors after being programmed. To prevent an approximate page from suffering too many errors that fail the application, we refresh each approximate page in six months. ApproxFTL provides the worst-case reliability guarantees under both programming errors and retention errors. Given that we cannot distinguish the error types, retention errors cannot be corrected at refresh time. We therefore upgrade approximate pages to precise pages when refreshing such pages.

E. Overhead Analysis

AproxFTL, while mitigating program disturbance and prolonging chip lifetime, introduces three types of overheads: hardware overhead, storage overhead, and firmware overhead. For the hardware overhead, we need the hardware support to enable approximate write. Similar approaches have been adopted in [9] with reasonable overhead.

For the storage overhead, we need three flags: (1) a 2-bit flag to indicate the page allocation pattern; (2) a 1-bit flag to indicate the type of the first page when the last time the block was used as $B^{used}_{approxPhase2}$ or B^{active}_{chb}. This flag is to assist intra-block wear leveling; Note, these three bits help to identify the block type of the to-be-programmed block. We choose reduced erase voltage to erase a block if all its pages are programmed as approximate pages, and choose normal erase voltage otherwise. (3) a 3-bit flag to indicate the maximal time that an approximate page in the block has been written using approximate write. This flag is to assist approximate page promotion. The seven bits are stored at the OOB (out-of-band) region of the first page for each block, which has negligible overhead for modern flash chips.

For the firmware overhead, we demand five active blocks, i.e., the integration of multi-streamed technology if it is not embedded already. Kang *et al.* showed that the overhead is negligible [39].

V. Performance Evaluation

In this section, we present the experimental methodology, evaluate the effectiveness of ApproxFTL, and analyze the experimental results with comparison to the state-of-the-art.

A. Experiment Setup

To evaluate the effectiveness of our proposed ApproxFTL, we implemented it in an event-driven SSD simulator. We simulated a 128GB 8-layer 3D NAND flash-based SSD [14], where the pages from one block are spread across all 8 layers, each layer has 64 pages from the block, and each page is of 8KB. The garbage collection (GC) is triggered when the number of free blocks goes below 10% of the total number of blocks. The GC operations are executed in the background in order to minimize the influence on the foreground requests. The key NAND flash parameters used in our simulation are listed in Table I.

In the baseline, the FTL adopts sequential write order and program all pages using precise writes. A programmed page, even after two disturb operations from its neighboring pages in Y and Z directions, respectively, has 10^{-8} or better RBER. The latency for read, program and erase operation is 45 *us*, 700 *us* and 3.5 *ms*, respectively.

TABLE I
Key Parameters of NAND Flash Memory

Flash Organization
density = 128 GB, total number of layers = 8, page size = 8 KB, page per block = 512, total blocks = 32768, GC threshold = 10%, over-provisioning ratio = 7%, page read = 45 *us*, page write = 700 *us*, block erase = 3.5 *ms*
Electrical Feature
DRAM R/W current = 125 mA, Flash R/W/E current = 25 mA, supply voltage = 3.3 V

The workloads. We used two types of workloads in the evaluation. First, the raw error-tolerance I/O traces is collected. In addition, we plugged in errors to random locations of the approximate data and evaluated the quality degradation of the output.

When running four scientific and numerical computing applications on a Centos-7 Linux box — *fft*, *smm*, *lu* and *sor*, we adapt the annotated type system, EnerJ [9], to distinguish between approximate and precise data types, and at

the same time we use the *blktrace* tool to capture the block level I/O traces. With type annotations, we can calculate the proportion of approximate non-critical data. Then approximate data types can be distinguished from the collected I/O traces. When application names are not these scientific computing applications in the I/O traces, e.g. swapper, its requests are regarded as traditional precise data; Data requests with the same application names are isolated the precise portion from the approximate portion in the equal proportion for simplicity. We also collected the I/O traces for one video and one image processing application treated as the same way — *img* and *vid*.

Output quality metrics. To evaluate the quality of the output, we adopted application-specific metrics, the same as these in previous works [9]. Since previous six I/O traces do not have the detailed data content for each request, public data sets are used for output quality evaluation. In particular, *dt* uses a decision tree to predict the context in the data set "sensorlog". *Svm*, which is a supervised learning model based on a support vector machine, and *ann* which is another trained classifier based on a feed-forward neural network, analyze the "pendigits" data set. And *knn*, k-nearest neighbor algorithm is used to classify multi-class subgenus of "Iris" data set. The quality metric in these classifiers is the classification accuracy on test datasets relative to the real classification results, which is consistent with previous works [9].

Performance metrics. We evaluated ApproxFTL using five metrics, average read response time, average write response time, wearing, total energy consumption, write amplification, and quality metric. The average read/write response time is a good metric for estimating ApproxFTL performance. And, the efficacy of the proposed scheme is also evaluated based the wearing for each FTL, which specifies the lifetime improvement under different FTLs. i.e., the wearing is more severe, then the lifetime of flash memory will be shorter. The total energy consumption measure is used to evaluate the impact of FTLs on energy consumption. All operations in flash-based SSD consume energy. For example, given that the time consumption of one flash erase operation is 3.5 *ms*, whereas the supplying voltage is 3.3 V and the erase current is 25 mA as listed in Table 1, one erase operation in flash consumes 288.75 μJ. The write amplification metric is used to evaluate the impact of GC on internal writes. The quality metric is evaluated based on quality degradation to demonstrate the effect of approximate storage on application-specific quality loss.

Schemes for comparison. In the experiments, we implemented and compared the following schemes.

- `Baseline`. This is the scheme that implements the traditional page-mapping FTL [51]. The baseline adopts sequential write order when programming data in the active block, and programs all the pages using traditional precise writes, i.e. t_PROG=700, t_R=45, t_BERS=3500 [52]. It does not exploit the error resilience in modern applications.
- `Delta_b`. This is the scheme that implements the simple approximate write [9]. It uses a larger ΔV_{pp} that is 1.5× of the baseline, and assumes page writes from upper level are tagged so the FTL knows if it needs approximate write or precise write. The RBER jumps to 7.2×10^{-4} from the baseline.
- `Vth_b`. This is the scheme that implements our V_{max_th}-reduced approximate write. The V_{max_th} is around 62% of the baseline. The RBER jumps to 7.2×10^{-4} from the baseline.
- `ApproxFTL`. This scheme is built on top of `Vth_b`. It integrates three enhancements in FTL to maximize disturbance mitigation.

B. Write Response Time Comparison

Fig. 5 compares the write response time from different schemes. The results were normalized to `Baseline`. From the figure, `ApproxFTL` reduces the average write response time by 45.64% over `Baseline`. The write performance improvement comes from: (i) Due to the reduction of ISPP steps in `ApproxFTL`, `Delta_b` and `Vth_b`, the response time of approximate write is reduced; (ii) Due to reduced disturbance, programming a precise page takes less time if it is mapped to $B^{active}_{approxPhase2}$ and B^{active}_{chb}. In summary, `ApproxFTL` achieves the largest reduction, i.e., 46.67%, in *vid*, and the smallest reduction, i.e., 5.15%, in *smm*.

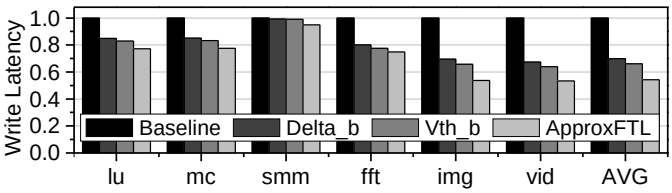


Fig. 5. The normalized write performance.

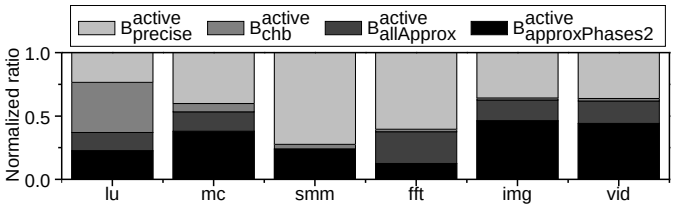


Fig. 6. The cumulative distribution of block instances in ApproxFTL.

To fully understand the write performance improvement in `ApproxFTL`, Fig. 6 reports the cumulative distribution of different active block instances. We ignore $B^{active}_{approxPhase1}$ as it is an intermediate state, which gets converted to either $B^{active}_{approxPhase2}$ or $B^{active}_{allApprox}$. Since writing pages in $B^{active}_{precise}$ is the same as the baseline, the more the pages are mapped to this type of active blocks, e.g., *smm*, the smaller improvement the write performance has.

B^{active}_{chb} has a small percentage in all applications except *lu*. This is because two-phase interleaving (i.e., $B^{active}_{approxPhases2}$) has better performance and thus we prioritize two-phase interleaving. Since all pages in $B^{active}_{allApprox}$ are written using fast approximate-write, a large percentage of $B^{active}_{allApprox}$, e.g, *fft*, tends to have better performance.

C. Read Response Time Comparison

Fig. 7 compares the read response time from different schemes, with the results normalized to `Baseline`. From the figure, `ApproxFTL` achieves 41.38% read response time reduction compared to `Baseline`. The improvement comes mainly from the reduction in read waiting time — approximate writes are faster so that the read operations wait shorter amount of time before being serviced.

The workload *mc* has negligible improvement because it only has a small read over write ratio, which means reducing write time has little impact on read waiting time.

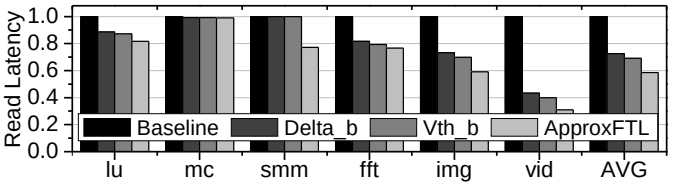


Fig. 7. The normalized read performance.

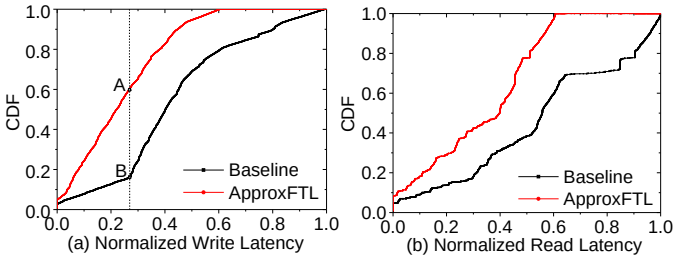


Fig. 8. The CDF of request latency.

Fig. 8 compares the cumulative distribution of different response time in servicing write and read requests, respectively, from workload *img* when adopting `ApproxFTL` and `Baseline`. We normalized the results to the worse write and read response time in `Baseline`, respectively. From the figure, more requests are serviced quicker in `ApproxFTL` than they are in `Baseline`. For example, the largest write response time in `ApproxFTL` is about 60.50% of that in `Baseline`.

Given an x-axis value, e.g., 0.269 or 26.9% of the longest latency in `Baseline`, we observed that about 61% (dot A) and 17% (dot B) of write requests are serviced with quicker than this time when using `ApproxFTL` and `Baseline`, respectively.

D. Energy Consumption Comparison

We next compared the energy consumption when adopting different schemes. We adopted the energy model from Hu *et al.* [53] to calculate the total energy consumption of the flash-based SSD and a typical 128MB DRAM buffer inside the flash module. We model all three basic operations, read/write/erase in flash. Fig. 9 summarizes the normalized results over `Baseline`. From the figure, on average, `ApproxFTL` achieves 24.5%, 10.3% and 8.2% improvements over `Baseline`, `Delta_b` and `Vth_b`, respectively.

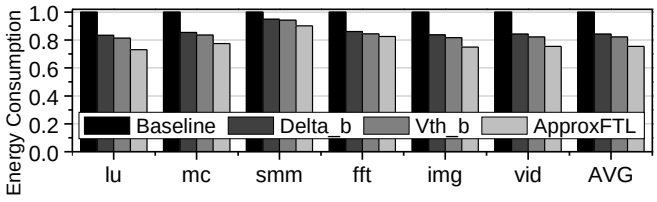


Fig. 9. Energy consumption comparison.

The large energy consumption reduction comes from two sources: (1) page reads and writes in `ApproxFTL` have shorter access latencies than they are in other schemes, and thus consume less energy; (2) `Vth_b` and `ApproxFTL` reduce the maximal threshold voltage when programming approximate pages such that each write consumes less energy than that in `Baseline`.

`ApproxFTL` achieves the greatest energy consumption reduction over `Baseline` in *lu*, i.e., 26.8%, and the smallest in *smm*, i.e., 9.9%.

E. Lifetime Comparison

We next compared the chip lifetime when adopting different schemes and summarized the results in Fig. 10. The results were normalized to `Baseline`. On average, `ApproxFTL` achieves about 5.75% chip lifetime improvement over the baseline.

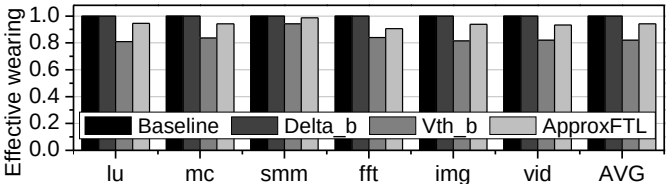


Fig. 10. Comparing chip lifetime.

`ApproxFTL` extends chip lifetime mainly because it uses a reduced erase voltage when erasing the blocks that only save approximate pages. The more $B^{used}_{approxPhase2}$ blocks the workload has, the large the lifetime improvement is. Since all approximate pages in the `Vth_b` scheme are in separate blocks, it shows slightly better lifetime improvement than `ApproxFTL`.

F. Impact on Write Amplification

Approximate and precise data may be mixed in one block, e.g. B^{active}_{chb} or $B^{active}_{approxPhases2}$, we should evaluate the impact of the proposed FTL on the garbage collection performance. In this experiment the flash memory storage will be filled with uniform random workloads before each simulation, so that GC and write amplification may be triggered for evaluation, and we will provide evaluation on how different writing schemes affect the write amplification of the flash storage.

Figure 11 compares the GC-induced write amplification in different schemes. `Delta_b` and `Vth_b` does not take hot/cold separation into consideration, and thus have the comparable write amplification quality as that in `baseline`.

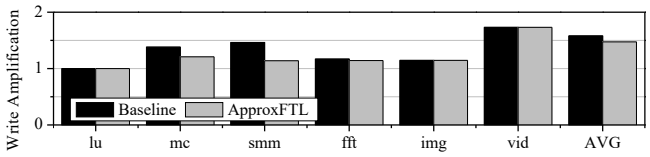


Fig. 11. Write amplification for each benchmark.

From the figure, `ApproxFTL` reduce the average write amplification by 6.64% over `baseline`. The write amplification reduction comes from: (1) Approximate data in scientific computing applications are more likely to be hot [54], and hot precise data will be allocated to B^{active}_{chb} or $B^{active}_{approxPhases2}$ blocks, thus we can expect that the victim hot GC block will be filled with invalid pages soon and it can be reclaimed with low overhead. (2) Separating hot write data into the hot pool in *img* and *vid* will largely reduce the overhead of garbage collection.

G. Impact on Quality

In the last experiment, we evaluated the quality of the output. Due to the limitation of our simulator, we can only plug in error simulating approximate-write introduced data inaccuracy in a subset of applications.

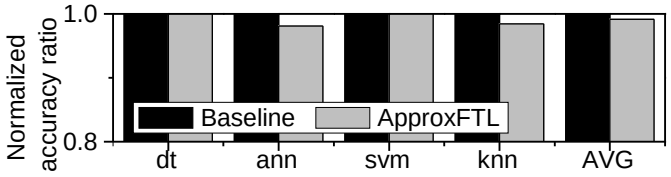


Fig. 12. Quality degradation for each benchmark.

Fig. 12 summarizes the normalized output quality from `ApproxFTL`. We used application-specific quality metrics that are widely in approximate computing research. `Delta_b` and `Vth_b` use the same RBER in relaxing approximate writes, and thus have the comparable output quality as that in `ApproxFTL`.

From the figure, we observed negligible reduction of output quality. There are three reasons: (1) Occasional errors for non-critical data show negligible impact on quality loss, since these applications have the inherent error resilience. (2) In implementing the approximate write, ApproxFTL chooses conservative parameters such that approximate pages accumulates small number of errors, even we map precise pages to their neighbors. (3) We employ approximate page promotion policy so that the system prevents approximate pages from unbounded error-prone approximate writes.

VI. Conclusion

In this paper, we proposed ApproxFTL, an approximate-write aware FTL design for 3D NAND flash memory storage systems. By reducing the maximal threshold voltage, we implement approximate write with significantly program disturbance reduction to neighboring pages.

Based on the tight correlation between reliability and performance, we devised pattern-guide page allocation design that prioritizes the interleaving of approximate and precise pages in LPN-to-PPN mapping. We then enhance wear leveling and garbage collection policies in FTL to take advantage of the mitigated program disturbance in approximate write. Our experimental results show that ApproxFTL, while preserving high data quality, improves the read and write response time of flash accesses by 41.38% and 45.64% on average, respectively, and extends the lifetime of 3D flash based SSDs by 5.75% when comparing to the state-of-the-art.

Acknowledgment

The authors would like to thank the anonymous reviewers for their detailed and thoughtful feedback which improved the quality of this paper significantly. We would also like to thank Adrian Sampson for providing approximate applications on persistent storage.

References

[1] Rino Micheloni. *3D Flash memories*. Springer, 2016.

[2] Teng-Hao Elton Yeh, Wei-Chen Chen, Tzu-Hsuan Bruce Hsu, Pei-Ying Penny Du, Chih-Chang Hsieh, Hang-Ting Lue, Yen-Hao Shih, Ya-Chin King, and Chih-Yuan Lu. Z-interference and z-disturbance in vertical gate-type 3-d nand. *IEEE Transactions on Electron Devices*, 63(3):1047–1053, 2016.

[3] Seiichi Aritome, Yoohyun Noh, Hyunseung Yoo, Eun Seok Choi, Han Soo Joo, Youngsoo Ahn, Byeongil Han, Sungjae Chung, Keonsoo Shim, Keunwoo Lee, et al. Advanced dc-sf cell technology for 3-d nand flash. *IEEE Transactions on Electron Devices*, 60(4):1327–1333, 2013.

[4] Chun-Hsiung Hung, Hang-Ting Lue, Kuo-Pin Chang, Chih-Ping Chen, Yi-Hsuan Hsiao, Shih-Hung Chen, Yen-Hao Shih, Kuang-Yeu Hsieh, Mars Yang, James Lee, et al. A highly scalable vertical gate (vg) 3d nand flash with robust program disturb immunity using a novel pn diode decoding structure. In *VLSI Technology (VLSIT), 2011 Symposium on*, pages 68–69. IEEE, 2011.

[5] Yi Wang, Zili Shao, Henry CB Chan, Luis Angel D Bathen, and Nikil D Dutt. A reliability enhanced address mapping strategy for three-dimensional (3-d) nand flash memory. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 22(11):2402–2410, 2014.

[6] Yu-Ming Chang, Yuan-Hao Chang, Tei-Wei Kuo, Yung-Chun Li, and Hsiang-Pang Li. Disturbance relaxation for 3d flash memory. *IEEE Transactions on Computers*, 65(5):1467–1483, 2016.

[7] Yu-Ming Chang, Yung-Chun Li, Yuan-Hao Chang, Tei-Wei Kuo, Chih-Chang Hsieh, and Hsiang-Pang Li. On relaxing page program disturbance over 3d mlc flash memory. In *Computer-Aided Design (ICCAD), 2015 IEEE/ACM International Conference on*, pages 479–486. IEEE, 2015.

[8] Hung-Sheng Chang, Yuan-Hao Chang, Tei-Wei Kuo, Yu-Ming Chang, and Hsiang-Pang Li. A disturbance-aware sub-block design to improve reliability of 3d mlc flash memory. In *Proceedings of the Eleventh IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, pages 1–7. ACM, 2016.

[9] Adrian Sampson, Jacob Nelson, Karin Strauss, and Luis Ceze. Approximate storage in solid-state memories. *ACM Transactions on Computer Systems (TOCS)*, 32(3):9:1–9:7, 2014.

[10] H Tanaka, M Kido, K Yahashi, M Oomura, R Katsumata, M Kito, Y Fukuzumi, M Sato, Y Nagata, Y Matsuoka, et al. Bit cost scalable technology with punch and plug process for ultra high density flash memory. In *VLSI Technology, 2007 IEEE Symposium on*, pages 14–15. IEEE, 2007.

[11] Dongku Kang, Woopyo Jeong, Chulbum Kim, Doo-Hyun Kim, Yong Sung Cho, Kyung-Tae Kang, Jinho Ryu, Kyung-Min Kang, Sungyeon Lee, Wandong Kim, et al. 256 gb 3 b/cell v-nand flash memory with 48 stacked wl layers. *IEEE Journal of Solid-State Circuits*, 52(1):210–217, 2017.

[12] Wonjoo Kim, Sangmoo Choi, Junghun Sung, Taehee Lee, Chulmin Park, Hyoungsoo Ko, Juhwan Jung, Inkyong Yoo, and Yoondong Park. Multi-layered vertical gate nand flash overcoming stacking limit for terabit density storage. In *VLSI Technology, 2009 Symposium on*, pages 188–189. IEEE, 2009.

[13] Hang-Ting Lue, Shih-Hung Chen, Yen-Hao Shih, Kuang-Yeu Hsieh, and Chih-Yuan Lu. Overview of 3d nand flash and progress of vertical gate (vg) architecture. In *Solid-State and Integrated Circuit Technology (ICSICT), 2012 IEEE 11th International Conference on*, pages 1–4. IEEE, 2012.

[14] Chih-Chang Hsieh, Hang-Ting Lue, Yung Chun Li, Kuo-Ping Chang, Hsing Chen Lu, Hsiang-Pang Li, Wei-Chen Chen, Yi-Hsuan Hsiao, Shuo-Nan Hung, Ti-Wen Chen, et al. Study of the interference and disturb mechanisms of split-page 3d vertical gate (vg) nand flash and optimized programming algorithms for multi-level cell (mlc) storage. In *VLSI Technology (VLSIT), 2013 Symposium on*, pages T156–T157. IEEE, 2013.

[15] Guoyu Wang, Hongsheng Zhang, Mingying Lu, Chao Zhang, Tao Jiang, and Guangyu Guo. Low-cost low-power asic solution for both dab+ and dab audio decoding. *IEEE transactions on very large scale integration (VLSI) systems*, 22(4):913–921, 2014.

[16] Hung-Sheng Chang, Yuan-Hao Chang, Tei-Wei Kuo, and Hsiang-Pang Li. A light-weighted software-controlled cache for pcm-based main memory systems. In *Computer-Aided Design (ICCAD), 2015 IEEE/ACM International Conference on*, pages 22–29. IEEE, 2015.

[17] Yu Cai, Onur Mutlu, Erich F Haratsch, and Ken Mai. Program interference in mlc nand flash memory: Characterization, modeling, and mitigation. In *Computer Design (ICCD), 2013 IEEE 31st International Conference on*, pages 123–130. IEEE, 2013.

[18] Jae-Duk Lee, Sung-Hoi Hur, and Jung-Dal Choi. Effects of floating-gate interference on nand flash memory cell operation. *IEEE Electron Device Letters*, 23(5):264–266, 2002.

[19] Yu Cai, Erich F Haratsch, Onur Mutlu, and Ken Mai. Error patterns in mlc nand flash memory: Measurement, characterization, and analysis. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 521–526. EDA Consortium, 2012.

[20] Yu Cai, Saugata Ghose, Yixin Luo, Ken Mai, Onur Mutlu, and Erich F Haratsch. Vulnerabilities in mlc nand flash memory programming: experimental analysis, exploits, and mitigation techniques. In *High Performance Computer Architecture (HPCA), 2017 IEEE International Symposium on*, pages 49–60. IEEE, 2017.

[21] Song Liu, Karthik Pattabiraman, Thomas Moscibroda, and Benjamin G Zorn. Flikker: saving dram refresh-power through critical data partitioning. *ACM SIGPLAN Notices*, 47(4):213–224, 2012.

[22] Adrian Sampson, Werner Dietl, Emily Fortuna, Danushen Gnanapragasam, Luis Ceze, and Dan Grossman. Enerj: Approximate data types for safe and general low-power computation. In *ACM SIGPLAN Notices*, volume 46, pages 164–174. ACM, 2011.

[23] Hadi Esmaeilzadeh, Adrian Sampson, Luis Ceze, and Doug Burger. Architecture support for disciplined approximate programming. In *ACM SIGPLAN Notices*, volume 47, pages 301–312. ACM, 2012.

[24] Qing Guo, Karin Strauss, Luis Ceze, and Henrique S Malvar. High-density image storage using approximate memory cells. In *ACM SIGPLAN Notices*, volume 51, pages 413–426. ACM, 2016.

[25] Djordje Jevdjic, Karin Strauss, Luis Ceze, and Henrique S Malvar. Approximate storage of compressed and encrypted videos. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 361–373. ACM, 2017.

[26] Daya S Khudia, Babak Zamirai, Mehrzad Samadi, and Scott Mahlke. Rumba: An online quality management system for approximate computing. In *Computer Architecture (ISCA), 2015 ACM/IEEE 42nd Annual International Symposium on*, pages 554–566. IEEE, 2015.

[27] Hyun-Seob Lee, Hyun-Sik Yun, and Dong-Ho Lee. Hftl: hybrid flash translation layer based on hot data identification for flash memory. *IEEE Transactions on Consumer Electronics*, 55(4):2005–2011, 2009.

[28] Mong-Ling Chiao and Da-Wei Chang. Rose: A novel flash translation layer for nand flash memory based on hybrid address translation. *IEEE Transactions on Computers*, 60(6):753–766, 2011.

[29] Sungjin Lee, Jihoon Park, Kermin Fleming, Jihong Kim, et al. Improving performance and lifetime of solid-state drives using hardware-accelerated compression. *IEEE Transactions on Consumer Electronics*, 57(4):1732–1739, 2011.

[30] Feng Chen, Tian Luo, and Xiaodong Zhang. Caftl: A content-aware flash translation layer enhancing the lifespan of flash memory based solid state drives. In *FAST*, volume 11, pages 77–90, 2011.

[31] Aayush Gupta, Youngjae Kim, and Bhuvan Urgaonkar. *DFTL: a flash translation layer employing demand-based selective caching of page-level address mappings*, volume 44. ACM, 2009.

[32] Xin Xu and H Howie Huang. Exploring data-level error tolerance in high-performance solid-state drives. *IEEE Transactions on Reliability*, 64(1):15–30, 2015.

[33] Jaeyong Jeong, Sangwook Shane Hahn, Sungjin Lee, and Jihong Kim. Improving nand endurance by dynamic program and erase scaling. In *HotStorage*, pages 1–5, 2013.

[34] Liang Shi, Kaijie Wu, Mengying Zhao, Chun Jason Xue, Duo Liu, and Edwin H-M Sha. Retention trimming for lifetime improvement of flash memory storage systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 35(1):58–71, 2016.

[35] Dae-Seok Byeon, Sung-Soo Lee, Young-Ho Lim, Jin-Sung Park, Wook-Kee Han, Pan-Suk Kwak, Dong-Hwan Kim, Dong-Hyuk Chae, Seung-Hyun Moon, Seung-Jae Lee, et al. An 8 gb multi-level nand flash memory with 63 nm sti cmos process technology. In *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International*, pages 46–47. IEEE, 2005.

[36] Yangyang Pan, Guiqiang Dong, and Tong Zhang. Exploiting memory device wear-out dynamics to improve nand flash memory system performance. In *FAST*, volume 11, pages 245–258, 2011.

[37] Neal Mielke, Todd Marquart, Ning Wu, Jeff Kessenich, Hanmant Belgal, Eric Schares, Falgun Trivedi, Evan Goodness, and Leland R Nevill. Bit error rate in nand flash memories. In *Reliability Physics Symposium, 2008. IRPS 2008. IEEE International*, pages 9–19. IEEE, 2008.

[38] JEDEC Standard. Stress-test-driven qualification of integrated circuits. *JEDEC Solid State Technology Association. JEDEC Standard*, (47F):1–26, 2007.

[39] Jeong-Uk Kang, Jeeseok Hyun, Hyunjoo Maeng, and Sangyeun Cho. The multi-streamed solid-state drive. In *HotStorage*, pages 1–5, 2014.

[40] Laura M Grupp, John D Davis, and Steven Swanson. The harey tortoise: Managing heterogeneous write performance in ssds. In *USENIX Annual Technical Conference*, pages 79–90, 2013.

[41] Liang Shi, Yejia Di, Mengying Zhao, Chun Jason Xue, Kaijie Wu, and Edwin H-M Sha. Exploiting process variation for write performance improvement on nand flash memory storage systems. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(1):334–337, 2016.

[42] Gala Yadgar, Eitan Yaakobi, and Assaf Schuster. Write once, get 50% free: Saving ssd erase costs using wom codes. In *FAST*, pages 257–271, 2015.

[43] Soojun Im and Dongkun Shin. Comboftl: Improving performance and lifespan of mlc flash memory using slc flash buffer. *Journal of Systems Architecture*, 56(12):641–653, 2010.

[44] Li-Pin Chang. Hybrid solid-state disks: combining heterogeneous nand flash in large ssds. In *Design Automation Conference, 2008. ASPDAC 2008. Asia and South Pacific*, pages 428–433. IEEE, 2008.

[45] Xavier Jimenez, David Novo, and Paolo Ienne. Wear unleveling: improving nand flash lifetime by balancing page endurance. In *FAST*, volume 14, pages 47–59, 2014.

[46] Saher Odeh and Yuval Cassuto. Nand flash architectures reducing write amplification through multi-write codes. In *Mass Storage Systems and Technologies (MSST), 2014 30th Symposium on*, pages 1–10. IEEE, 2014.

[47] Yixin Luo, Yu Cai, Saugata Ghose, Jongmoo Choi, and Onur Mutlu. Warm: Improving nand flash memory lifetime with write-hotness aware retention management. In *Mass Storage Systems and Technologies (MSST), 2015 31st Symposium on*, pages 1–14. IEEE, 2015.

[48] Renhai Chen, Yi Wang, Duo Liu, Zili Shao, and Song Jiang. Heating dispersal for self-healing nand flash memory. *IEEE Transactions on Computers*, 66(2):361–367, 2017.

[49] Renhai Chen, Yi Wang, Jingtong Hu, Duo Liu, Zili Shao, and Yong Guan. Image-content-aware i/o optimization for mobile virtualization. *ACM Transactions on Embedded Computing Systems (TECS)*, 16(1):12:1–12:24, 2016.

[50] Renhai Chen, Yi Wang, and Zili Shao. Dheating: Dispersed heating repair for self-healing nand flash memory. In *Proceedings of the Ninth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, pages 1–10. IEEE Press, 2013.

[51] Yang Hu, Hong Jiang, Dan Feng, Lei Tian, Hao Luo, and Shuping Zhang. Performance impact and interplay of ssd parallelism through advanced commands, allocation strategy and data granularity. In *Proceedings of the international conference on Supercomputing*, pages 96–107. ACM, 2011.

[52] Jae-Woo Im, Woo-Pyo Jeong, Doo-Hyun Kim, Sang-Wan Nam, Dong-Kyo Shim, Myung-Hoon Choi, Hyun-Jun Yoon, Dae-Han Kim, You-Se Kim, Hyun-Wook Park, et al. 7.2 a 128gb 3b/cell v-nand flash memory with 1gb/s i/o rate. In *Solid-State Circuits Conference-(ISSCC), 2015 IEEE International*, pages 1–3. IEEE, 2015.

[53] Yang Hu, Hong Jiang, Dan Feng, Lei Tian, Shuping Zhang, Jingning Liu, Wei Tong, Yi Qin, and Liuzheng Wang. Achieving page-mapping ftl performance at block-mapping ftl cost by hiding address translation.

In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, pages 1–12. IEEE, 2010.

[54] Vinay K Chippa, Srimat T Chakradhar, Kaushik Roy, and Anand Raghunathan. Analysis and characterization of inherent application resilience for approximate computing. In *Proceedings of the 50th Annual Design Automation Conference*, pages 113–121. ACM, 2013.

Jinhua Cui received the B.S. degree from Southwest University, Chongqing, China, in 2012. She has also been a visiting student in the Computer Science Department at University of Pittsburgh, Pittsburgh, PA, USA, during the period between Aug. 2016 and Aug. 2017, co-advised by Professor Youtao Zhang and Professor Jun Yang.

She is currently a Ph.D. candidate of Xi'an Jiaotong University, Xi'an, Shaanxi, China, under the guidance of Professor Weiguo Wu. Her current research interests include NAND flash memory based storage system, approximate storage, big data, cloud computing, HDFS, and database index.

Youtao Zhang received the Ph.D. degree in computer science from the University of Arizona, Tucson, AZ, USA, in 2002, and the B.S. and M.E. degrees from Nanjing University, Nanjing, China, in 1993 and 1996, respectively.

He is currently an Associate Professor of Computer Science, University of Pittsburgh, Pittsburgh, PA, USA. His current research interests include computer architecture, program analysis, and optimization. Prof. Zhang was the recipient of the U.S. National Science Foundation Career Award in 2005. He is a member of ACM. His current research interests include onchip interconnection, architectural support for security, new memory technologies, and networks-on-chip.

Liang Shi received the B.S. degrees in computer science from the Xi'an University of Post and Telecommunication, Xi'an, Shaanxi, China, in 2008, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2013.

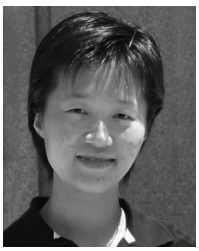
He is currently an Associate Professor with the College of Computer Science, Chongqing University, Chongging, China. His current research interests include flash memory, embedded systems, and emerging nonvolatile memory technology.

Chun Jason Xue received B.S. degree in Computer Science and Engineering from University of Texas at Arlington in May 1997, and M.S. and Ph.D. degree in computer Science from University of Texas at Dallas, in Dec 2002 and May 2007, respectively.

He is now an Assistant Professor in the Department of Computer Science at the City University of Hong Kong. His research interests include memory and parallelism optimization for embedded systems, software/hardware co-design, real time systems and computer security.

Weiguo Wu received his B.S., M.S. and Ph.D. degrees in computer science, all from Xi'an Jiaotong University, Shaanxi, China, in 1986, 1993 and 2006, respectively.

He is currently a Professor with the school of Electronic and Information Engineering, Xi'an Jiaotong University, Shaanxi, China. He is also the deputy director of the Neo Computer Institute in Xi'an Jiaotong University, a senior member of Chinese Computer Federation, a standing committee member of High Performance Computing Clusters in Chinese Computer Federation, and a standing committee member of Microcomputer (Embedded System) in Chinese Computer Federation, the director of Shaanxi Computer Federation. His research interests include high performance computing, computer network, embedded system, VHDL, cloud computing, and flash memory.



Jun Yang received the B.S. degree in computer science from Nanjing University, China, in 1995, the M.A. degree in mathematical sciences from Worcester Polytechnic Institute, Massachusetts, in 1997, the PhD degree in computer science from the University of Arizona in 2002.

She is a Professor with the Electrical and Computer Engineering Department, University of Pittsburgh, Pennsylvania. She is the recipient of US NSF Career Award in 2008. Her research interests include low power, and temperature-aware micro-architecture designs, new memory technologies, and networks-on-chip.