Improving Performance for Flash-Based Storage Systems through GC-Aware Cache Management

Suzhen Wu, *Member, IEEE*, Bo Mao, *Member, IEEE*, Yanping Lin, and Hong Jiang, *Fellow, IEEE*

Abstract—Flash-based SSDs have been extensively deployed in modern storage systems to satisfy the increasing demand of storage performance and energy efficiency. However, Garbage Collection (GC) is an important performance concern for flash-based SSDs, because it tends to disrupt the normal operations of an SSD. This problem continues to plague flash-based storage systems, particularly in the high performance computing and enterprise environment. An important root cause for this problem, as revealed by previous studies, is the serious contention for the flash resources and the severe mutually adversary interference between the user I/O requests and GC-induced I/O requests. The on-board buffer cache within SSDs serves to play an essential role in smoothing the gap between the upper-level applications and the lower-level flash chips and alleviating this problem to some extend. Nevertheless, the existing cache replacement algorithms are well optimized to reduce the miss rate of the buffer cache by reducing the I/O traffic to the flash chips as much as possible, but without considering the GC operations within the flash chips. Consequently, they fail to address the root cause of the problem and thus are far from being sufficient and effective in reducing the expensive I/O traffic to the flash chips that are in the GC state. To address this important performance issue in flash-based storage systems, particularly in the HPC and enterprise environment, we propose a Garbage Collection aware Replacement policy, called GCaR, to improve the performance of flash-based SSDs. The basic idea is to give higher priority to caching the data blocks belonging to the flash chips that are in the GC state. This substantially lessens the contentions between the user I/O operations and the GC-induced I/O operations. To verify the effectiveness of GCaR, we have integrated it into the SSD extended Disksim simulator. The simulation results show that GCaR can significantly improve the storage performance by up to 40.7 percent in terms of the average response times.

Index Terms—Flash storage, solid state drive, garbage collection, cache management, performance evaluation

1 Introduction

The development of multi-core and GPU processors in computer systems has further widened the performance gap between processors and storage devices. Hard Disk Drives (HDD) have been the primary storage deices in large-scale storage systems for a few decades now. The performance of HDDs has been improved rather slowly relative to other layers/components of the computer systems stack, making them the performance wall of storage systems. With advancements in the semi-conductor technology, flash-based SSDs have emerged as an appealing alternative or supplement to HDDs owing to their many attractive features such as light weight, high random-access performance and shock resistance. As a result, they have received a great deal of attention from both academia and industry. Besides the deployment on mobile devices and desktop/laptop PCs, they have been widely deployed in the high performance computing and enterprise environments [1], [2], [3], [4], [5], [6].

Flash-based SSDs consist of multiple flash chips with each containing a large amount of flash blocks. Each block consists of a fixed number of pages [1], [7]. Besides the normal read and write operations, block erase is also a frequent operation within flash-based SSDs because the pages can only be written once before the entire block is erased. Erase operations in flash memory are nearly an order of magnitude slower than write operations. Therefore, flash-based SSDs use out-of-place writes instead of the in-place writes used on HDDs. To reclaim the invalid pages and to create free space for writes, SSDs use a Garbage Collection (GC) process [8]. The GC process is a time-consuming operation since it copies the valid pages in blocks into the free storage pool and then erases the blocks that do not store valid data. A block erase operation takes approximately 2 milliseconds [1]. Considering that the valid pages in the victim blocks need to be copied and then erased, the GC overhead can be quite significant.

Usually, GC operations can be executed when there is sufficient idle time (i.e., no incoming I/O requests to SSDs) to minimize the impact on the user performance. However, workloads in server-centric enterprise data centers and High Performance Computing (HPC) environments often have bursts of requests with very short inter-arrival times. These workloads do not exhibit sufficiently long idle times [9], [10] to accommodate GC operations. Examples of enterprise workloads that exhibit this behavior include online-transaction

- *S. Wu and Y. Lin are with the Computer Science Department of Xiamen University, Xiamen, Fujian 361005, China. E-mail: suzhen@xmu.edu.cn, linyapppp@foxmail.com.*
- *B. Mao is with the Software School of Xiamen University, Xiamen, Fujian 361005, China. E-mail: maobo@xmu.edu.cn.*
- *H. Jiang is with the Computer Science and Engineering Department, at the University of Texas at Arlington, Arlington, TX 76019. E-mail: hong.jiang@uta.edu.*

Manuscript received 29 Sept. 2016; revised 20 Mar. 2017; accepted 4 Apr. 2017. Date of publication 12 Apr. 2017; date of current version 13 Sept. 2017. Recommended for acceptance by Z. Chen.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPDS.2017.2692757

processing applications, such as OLTP and OLAP [2], [11]. Furthermore, it has been found that HPC file systems are routinely stressed with write requests of frequent and periodic checkpointing and journaling operations [12]. In a study of HPC I/O workload characterization of the Spider storage system at Oak Ridge National Laboratory, it was observed that the bandwidth distributions are heavily long-tailed [13]. Thus, in the real environment, most GC operations are triggered on-demand. Our preliminary experiment results, detailed in Section 2.1, and previous studies [8], [14] observed that the GC operations can significantly degrade the user performance for bursty write-dominant workloads. All these studies have implied or shown that GC is a significant performance concern for flash-based storage systems in enterprise and HPC environments.

A lot of studies have been proposed to address the GC issues within flash-based SSDs in the literature, including different Flash Translation Layer (FTL) designs [10], [15], write reduction by deduplication and compression [16], [17], [18], [19], [20], and sophisticated data layout schemes that exploit the workload characteristics and internal parallelism [14], [21], [22], [23]. However, since they only try to optimize the random write traffic to alleviate the GC-induced overhead, they all fail to consider the contention for resources and adversary interference between the user I/O traffic and the GC-induced I/O traffic. During an ongoing GC process of a flash chip, any incoming user I/O request to this chip will be delayed until the completion of the GC. Thus, the GC process can significantly impede the user I/O performance by increasing their read and write queuing delays.

On the other hand, in order to bridge the performance gap between the upper-level host memory and the lower-level flash storage, the flash-based SSDs are usually embedded with an on-board buffer cache to smooth the user I/O traffic [1], [24]. In other words, the on-board buffer cache has the capability of changing the user accesses seen by the flash chips by absorbing bursty traffic and leveraging access locality with an appropriate cache replacement policy. The existing cache replacement policies, such as LRU and LAMA [25], are well optimized to maximally reduce the buffer cache miss rate by reducing the I/O traffic to the backend devices. The variants of these policies designed to improve GC efficiency, such as BPLRU [26] and GC-ARM [27], only consider how to replace data blocks to improve the GC efficiency and are not aware of the underlying GC operations within flash chips. Thus, none of them takes the GC operations within flash chips into consideration, making them insufficient and ineffective in reducing the expensive I/O traffic to the flash chips that are in the GC state. GC state indicates that a flash chip is conducting garbage collection: coping valid data from a block and erasing the block to free up the flash space for subsequent write data. We argue that, to alleviate the aforementioned contention and interference between the user I/O traffic and GC-induced I/O traffic, the management of the on-board buffer cache should be made aware of the real-time GC activities within the flash chips.

In this paper, we propose a garbage collection aware cache management, named GCaR, to substantially reduce the contention and interference between the user I/O operations and the GC I/O operations. The basic idea is to give a higher priority to caching the blocks on the flash chips that are in GC state. When replacing or destaging a data block in the cache line, we will check whether the data block belongs to the flash chips that are in or will soon be in the GC state. If so, the data block will be kept in the buffer cache for a longer time until the GC completes. Otherwise, it will be replaced or destaged as usual. Different from the existing goal of the cache management aiming to reduce as many user I/O requests as possible, the objective of GCaR is to reduce as many "right" kind of I/O requests as possible to significantly alleviate the contention and interference between the user I/O operations and the GC-induced I/O operations. We have integrated GCaR into the SSD extended Disksim simulator and conducted extensive evaluations. The performance results show that GCaR can significantly improve the storage performance by reducing the average response time by up to 40.7 percent.

To the best of our knowledge, GCaR is the first cache management scheme that considers the GC activities within flash chips for flash-based SSDs. This work achieves the following contributions:

- From the experimental results, we observed that GC can significantly render the SSD performance variable and unpredictable. Thus, the miss penalties in the run-time GC-active flash chips and GC-inactive flash chips are different from the point of the cache management.
- We propose a GC-aware cache management scheme that is aware of the internal GC activities of the flash-based SSDs to to significantly alleviate the contention and interference between the user I/O operations and the GC-induced I/O operations.
- We conduct extensive evaluations with both benchmark and trace-driven experiments to show the effectiveness and efficiency of GCaR. The evaluation results show that GCaR can improve the performance by up to 40.7 percent in terms of the average response time. We also conduct experiments of GCaR on SSD-based RAIDs, with results demonstrating the good extendibility of GCaR.

The rest of this paper is organized as follows. Background and motivation are presented in Section 2. We describe the design details of GCaR in Section 3. The performance evaluation is presented in Section 4 and the related work is presented in Section 5. We conclude this paper in Section 6.

2 Background and Motivation

In this section, we first present the key performance characteristics of SSDs most relevant to this research. Then we discuss factors affecting and affected by cache efficiency to motivate our proposed GC-aware cache management for flash-based SSDs.

2.1 SSD Basics

Different from mechanical HDDs, flash-based SSDs are made of silicon memory chips and do not have moving parts (i.e., mechanical positioning parts). Fig. 1 illustrates a logical overview of a typical SSD with n independent

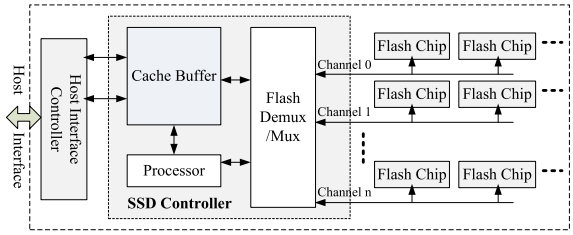

Fig. 1. Overview of a typical SSD showing its internal cache buffer and flash chips.

channels, each channel is shared by multiple flash chips. Despite of the high energy-efficiency and high random-read performance advantages of flash-based SSDs, they have the following two main unique non-HDD characteristics.

First, flash-based SSDs have asymmetric read-write-erase performance [1], [20], [28]. Writing to NAND flash is a multi-step process. In general, to write to cells that have existing data, one must first read the cells, which is followed by erasing the cells and then programming (writing) to the cells. In particular, SSDs suffer from the poor performance of small random-write requests. This leads to erase/program (writing over existing data) operations that are 2-3 orders of magnitude slower than reads. Furthermore, there are at least three types of NAND flash cells: Single-Level Cell (SLC), Multi-Level Cell (MLC) and Triple Level Cell (TLC). SLC can store a single bit of data. MLC and TLC can store two or three bits per cell. MLC can store at least twice the amount of data compared to SLC. But the read performance of MLC is twice as slow as SLC as could be expected and the write performance is over 3 times slower. To better understand the performance of the NAND flash chips, Table 1 shows the concrete examples of the read-write-erase performance asymmetry and the performance differences among SLC, MLC and TLC.

Second, GC is an important but expensive background process in flash-based SSDs. GC eliminates the need to perform erasure of the whole block prior to every write. It accumulates data marked for erase as "Garbage" and performs whole block erase as space reclamation in order to reuse the block. The required GC operations in SSD significantly affect the user I/O performance. The granularity of each read or write operation is a page (2KB-4KB) while that of an erase operation is a block (128 KB-256 KB). The execution time of an erase operation is one or two orders of magnitude more than that of a write or read operation respectively, as shown in Table 1. The reason is that, for flash-based SSDs, each block (128-256 KB) must be erased in advance before any part of it can be written, which is the characteristic

TABLE 1
Read/Program/Erase Times for SLC/MLC/TLC Chips [7]

Operations	SLC chips	MLC chips	TLC chips
Random Read (page)	25 us	50 us	80 us
Program (page write)	250 us	900 us	2.3 ms
Erase per block	2 ms	2 ms	10 ms

feature of flash chips and known as "*erase-before-write*". Due to the sheer size of a block, an erase operation typically takes milliseconds to complete. When the number of free blocks in an SSD is smaller than a predetermined threshold, the valid pages in the victim blocks (i.e., to be erased) must be copied to a different free block before the victim blocks are erased to create new free blocks (thus free pages), which is called garbage collection. The GC process can significantly degrade both read and write performance by increasing the queuing times of the user requests.

Fig. 2 shows the microscopic analysis of the average response times driven by the three realistic traces on the Intel DCS3700 200GB SSD. Before the evaluations, the SSD is filled with the written data. We can see that larger latencies are occurred due to the frequent GC operations. The large latencies are orders of magnitude than that in the normal state without GC operations. Moreover, previous studies also have similar findings [14] and shown that GC can render the SSD performance significantly variable and unpredictable in high performance computing and enterprise environments [4], [8], [29]. The Solid State Storage Initiative of SNIA has initiated a project named "Understanding SSD Performance Project" [30], which has found that, along with the performance evaluations, the performance of flash-based SSDs also degrades dramatically. All these studies have revealed that GC process has a significantly impact on the system performance.

2.2 Cache Efficiency

In storage systems, the buffer cache is widely used. The reason is that different storage devices, such as register, RAM, flash and disk, have different performance characteristics. The buffer cache is an effective way to bridge the performance gap between two neighboring layers of the storage stack that have storage devices with very different performance characteristics. The efficiency of the buffer cache can be assessed by the average memory access time that depends on the *Hit_Time*, *Miss_Rate* and *Miss_Penalty*. The Average Memory Access Time (AMAT) is represented by

$$AMAT = Hit\_Time + Miss\_Rate * Miss\_Penalty. \quad (1)$$

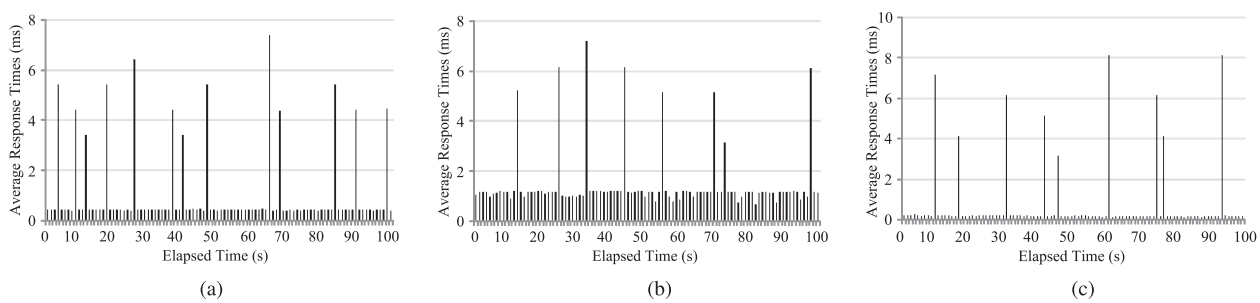

Fig. 2. The microscopic analysis of the average response times driven by the realistic traces.

The AMAT is depended on the three factors: the *Hit_Time*, the *Miss_Rate* and the *Miss_Penalty*. Each of them can be optimized to reduce the AMAT. As shown in Fig. 1, the on-board buffer cache is made up of the DRAM whose access time is consistent, meaning that the *Hit_Time* is consistent. The existing optimizations are trying to reduce the *Miss_Rate* as much as possible by reducing the number of the user I/O requests actually passed to the backend flash chips by exploiting access locality. They assume that the *Miss_Penalty* is also consistent because the direct access times from the flash chips are almost the same. While this assumption is valid for HDDs, it may not be fully validated for flash-based SSDs, as we discussed in Section 2.1. The access times of the flash chips can be quite different and varying depending on the presence or absence of the GC operations. The access time to the flash chip that is in the GC state will be much longer than that to one not in the GC state, by up to orders of magnitude. Thus we argue that *for the on-board buffer cache within flash-based SSDs, not only the*Miss_Rate, *but also the*Miss_Penalty*should be considered in the buffer cache management*.

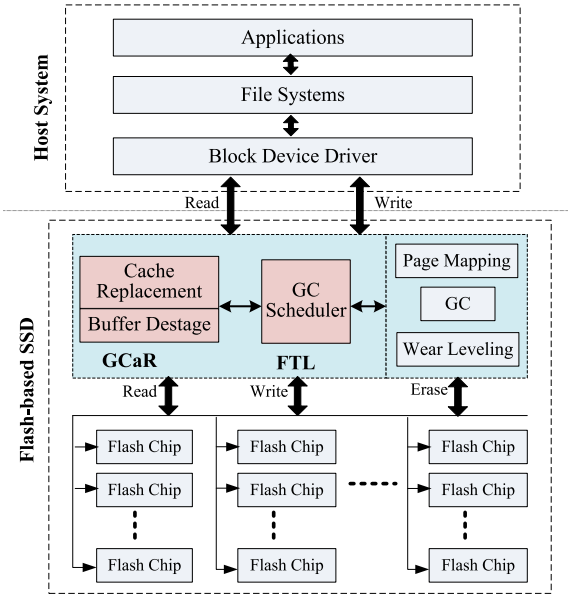

Fig. 3. Overview of GCaR on the I/O path.

2.3 Motivation

Reducing the miss rate is not the only way to improve the efficiency of the buffer cache. Reducing the miss penalty is also an important way in some conditions. Previous studies on the power-aware cache management scheme PB-LRU [31] and availability cache management schemes VDF [32] and Shaper [33] also found that under some conditions, the *Miss_Penalty* to the backend HDDs are different. For example, in PB-LRU, the access times on the active disks and sleep disks are different. In disk arrays with a faulty disk, the access times on the surviving disks and the faulty disk are also different. Thus, PB-LRU, VDF and Shaper take the *Miss_Penalty* as an important factor into their optimization considerations for the management of the buffer cache.

The access times on the flash chips with GC and without GC are significantly different for the on-board buffer cache within the flash-based SSDs. However, the existing cache replacement schemes, such as BPLRU [26] and GC-ARM [27], only consider how to replace the data blocks to improve the GC efficiency and are not aware of the underlying GC operations within the flash chips. Inspired by the previous studies on the power-aware cache management [31] and availability cache management [32], [33], we propose a GC-aware replacement scheme for buffer cache management, called GCaR, to improve the efficiency of the on-board buffer cache within the flash-based SSDs. By considering not only the *Miss_Rate*, but also the *Miss_Penalty*, the AMAT with GCaR will be shown to significantly improve.

3 GCaR

In this section, we first outline the main principles guiding the design of GCaR. Then we present a system overview of the GCaR, followed by a description of the the cache replacement and the buffer destaging algorithms in GCaR.

3.1 Design Principles

The design of GCaR aims to achieve high user performance, high GC efficiency and high portability, as explained below.

- *High User Performance*. Most of the existing on-board buffer cache management schemes are designed to reduce the miss rate by reducing the user I/O requests delivered to the flash chips, oblivious of the GC activities in these chips. Our GCaR buffer cache management scheme is designed to be made garbage collection-aware in that it not only considers the miss rate, but also considers the GC-dependent miss penalty. As a result, GCaR will be able to reduce the user average response time by keeping as many of the expensive user I/O requests in the memory as possible, thus avoiding the long waiting time on the flash chips that are in the GC state.
- *High GC Efficiency*. GCaR is designed to reduce the contention and interference between the user I/O requests and the GC-induced requests by keeping the interfering user requests in memory. This enables the GC operations to perform in the flash chips unimpeded to significantly improve the GC efficiency.
- *High Portability*. Most of the existing on-board buffer cache management schemes are designed to reduce the miss rate by reducing the user I/O requests delivered to the flash chips. Our GCaR buffer cache management strategy not only considers the miss rate, but also consider the GC-dependent miss penalty. We have evaluated and validated its portability by embedding it into LRU, CFLRU and BPLRU schemes to show its efficiency.

3.2 System Overview

Fig. 3 shows a system overview of our proposed GCaR within the flash translation layer (FTL) of a flash-based SSD on the system I/O path. The host consists of the applications, file system and block device driver. The SSD device receives the read and write requests from the block device driver through the host interface. The on-board buffer cache management will check whether the requested data is in the buffer cache. If so, they will be serviced by the on-board buffer cache. Otherwise they

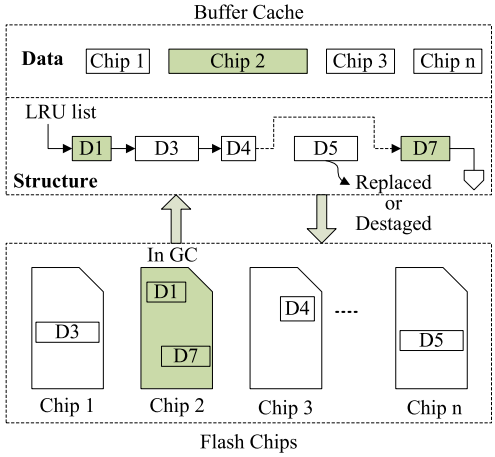

Fig. 4. The data structure in GCaR.

will be issued to the backend flash chips. It also must be noted that the GCaR scheme can also be extended to the buffer cache management for the Redundant Array of Independent SSDs (RAIS) system [34]. By replacing the flash chips with the SSDs in Fig. 3, the GCaR scheme can also alleviate the performance variability caused by the GC operations within each SSD in the RAIS system, as validated in Section 4.4.

GCaR has three main functional modules: the cache replacement module, the buffer destaging module and the GC scheduler module. The cache replacement and buffer destaging modules interact with the GC scheduler module to determine which data blocks should be replaced or destaged to reduce the miss penalty, as shown in Fig. 3. The GC scheduler module has access to the real-time information about whether a given flash chip is in the GC state or not by its interaction with the FTL. There are three operations that are issued from the FTL to the flash chips: read, write and erase. The GC process will invoke all the three operations within the involved flash chips, which significantly and adversely affects the user read and write performance. Thus, the miss penalty to a flash chip in the GC state will be waiting for a much longer time than that to one not in the GC state and thus it is very expensive.

GCaR not only exploits the locality of workloads, but also takes the miss penalty as an important design factor in the cache replacement and buffer destage, compared with the traditional buffer cache management schemes for flash-based SSDs. Fig. 4 shows the data structure in GCaR. In GCaR, all the data blocks in the buffer cache are associated with an LRU list to capture the workload locality. However, our scheme can also integrate with the data structures of schemes with other replacement algorithms, such as CFLRU [35] and LAMA [25]. Moreover, the conversion from the original cache algorithms to the GCaR-based algorithms will be smooth, because GCaR takes effect only when any of the flash chips are in the GC state. In other words, the buffer cache is managed by the original algorithms in the GC-free mode, and the GCaR policy becomes effective only when GC operations occur.

Fig. 4 shows an example of the LRU scheme with Chip 2 being in the GC state. Based on the LRU scheme, block D7 at the end of LRU list should be evicted from the buffer cache. However, by checking the logical-block-address to physical-block-address (LBA-PBA) mapping information, block D7 is in Chip 2 that is in the GC state, the GCaR scheme will thus keep it longer in the buffer cache and instead evict block D5 to free buffer space. Furthermore and importantly, when a flash chip is in the GC state, the buffer space allocated to that flash chip is also increased to further reduce the miss rate because of the significantly increased miss penalty in a chip in the GC state. In the following two sections, we will illustrate the cache replacement and buffer destaging schemes in GCaR in more detail.

3.3 Cache Replacement

In the traditional cache replacement algorithms for the on-board buffer cache within flash-based SSDs, they always assume that the miss penalty of each miss at the same level is of the same constant value, thus they by and large are designed to improve the hit rate on the cache by exploiting the request access locality. However, in flash-based SSDs with GC operations, the penalty of a miss to the missing data in the flash chips that are in the GC state are likely to be much more expensive than that of a miss to the flash chips not in the GC state. Therefore, from the aspect of an SSD device, the buffer cache performance should not be simply evaluated by the traditional metrics such as hit ratio or miss ratio. To address this issue, the cache replacement in GCaR takes not only the hit rate, but also the miss penalty as an important design parameter. A detailed description of the cache replacement algorithm in GCaR is given in Algorithm 1.

Algorithm 1. Cache Replacement Algorithm in GCaR

```
1:  Input: the user read requests R_1, R_2, . . .R_i, . . .
2:  procedure GCaR_LRU_Replace(R_i)
3:    if R_i is LRU_List then
4:      /*A cache hit case*/
5:      Return_from_cache(R_i)
6:    else
7:      /*A cache miss case*/
8:      if Cache is full then
9:        repeat Get bottom item in LRU_List: L_x
10:         if L_x belongs to the GC-flash-chip then
11:           x = x-1
12:         else
13:           Delete(LRU_List, L_x)
14:           Return_from_flash(R_i)
15:           Update(LRU_List, R_i)
16:           break
17:         end if
18:       until x == 0
19:     else
20:       Return_from_flash(R_i)
21:       Update(LRU_List, R_i)
22:     end if
23:   end if
24: end procedure
```

The GCaR scheme will check whether a read request hits the cache upon receiving the read request. If yes, the cached data will be returned. Otherwise, it will be fetched from the flash and the free cache space availability will be checked. If the cache is full, GCaR will replace data blocks to make free space for the new fetched data. The data blocks at the LRU

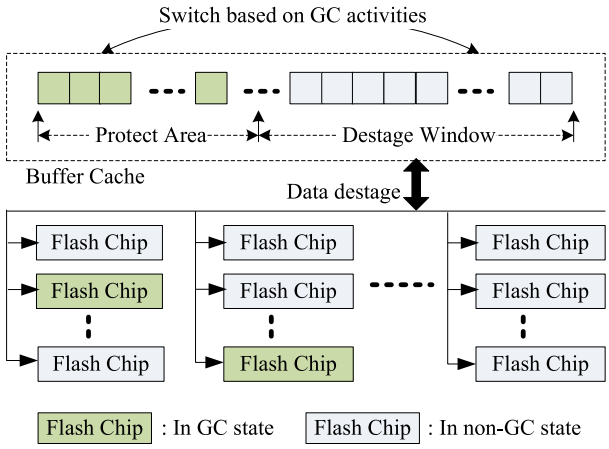

Fig. 5. The buffer destaging workflow in GCaR.

tail will be checked to make sure that the evicted data blocks do not belong to the flash chips that are in the GC state. Once determined, the new fetched data from flash will be kept in the cache and the LRU list is updated. If the cache has free space, the new fetched data from flash will be written to the cache directly and the LRU list is updated. After the requested data is fetched from the flash to the cache, it will be returned to the upper layer.

The miss penalty on the flash chips that are in the GC state is orders of magnitude larger than the miss penalty on the flash chips not in the GC state, thus the data blocks belonging to the former will be given a higher priority of being retained in cache in the GCaR scheme. When encountering a read miss and in need of replacing an existing data block in the cache, the data block from the tail of the LRU list not given the higher priority (i.e., not belonging the a chip in the GC state) will be selected as a victim data block. As shown in Fig. 4, since data block D7 belongs to the Chip 2 that is in the GC state, data block D5 belonging to Chip n (not in the GC state) will be evicted from the cache to free buffer space.

The GC operations are controlled and scheduled by the FTL within flash-based SSDs. The GC scheduler module in GCaR is responsible for getting the GC state information of each chip. Based on the GC state of a given flash chip, the GCaR can check whether a cache block should be replaced belongs to the flash chips that are in the GC state. In a real environment, the GC operations are triggered dynamically for each flash chips. Thus the priority of a data block in the cache also changes dynamically according to the underlying GC operations. Adding and updating a priority tag in the metadata for each of the cached data blocks can introduce a significant performance overhead. In our current design, the check for the GC state only happens when a data block needs to be replaced to avoid the metadata update overhead.

3.4 Buffer Destaging

Buffer plays an important role in improving the write performance by absorbing and filtering the write traffic from the upper level before going to the flash chips. For performance purposes, the buffer is usually set up with a write-back scheme. If the buffer space is unlimited, then the write performance will be the same as the memory speed. However, due to the cost and power issues, the buffer space is limited. The existing buffer management schemes concentrate on first merging random writes into sequential ones to improve both the write performance when destaging to the chips and the GC efficiency by reducing the GC-induced I/O traffic [26], [27]. However, these schemes, being unaware of the GC state of the involved chips, are only beneficial when writing to chips not in the GC state and are unable to improve the user write performance during GC operations.

Fig. 5 shows the buffer destaging workflow in the GCaR schemes. The data unit in the buffer cache is a page or a block determined by the different cache management schemes. The buffer destaging scheme in GCaR takes the GC operations into consideration by first detaging data blocks belonging to the flash chips that are not in the GC state. To make free space for the subsequent write data on the flash chips that are in the GC state, the GCaR scheme will allocate much more buffer space to temporally store all the write data blocks belonging to the flash chips that are in the GC state. Logically, the buffer cache is partitioned into the protect area and the destage area. The two areas are dynamically changed based on the GC activities within flash chips. Once a flash chip completes its GC process, the corresponding data blocks belonging to that flash chip will be migrated from the protect area to the destage area. By avoiding issuing the user write traffic to the flash chips that are in the GC state, both the GC efficiency and the user write performance can be improved.

Algorithm 2. Buffer Destage Algorithm in GCaR

```
1:  Input: the user write requests W_1, W_2, ...W_i, ...
2:  procedure GCaR_LRU_Destage(W_i)
3:    if W_i is LRU_List then
4:      /*A buffer hit case*/
5:      Write_to_buffer(W_i)
6:    else
7:      /*A buffer miss case*/
8:      if Buffer is full then
9:        repeat Get bottom item in LRU_List: L_x
10:         if L_x belongs to the GC-flash-chip then
11:           x = x-1
12:         else
13:           Write_to_flash(D_x)
14:           Delete(LRU_List, L_x)
15:           Write_to_buffer(W_i)
16:           Update(LRU_List, W_i)
17:           break
18:         end if
19:       until x == 0
20:     else
21:       Write_to_buffer(W_i)
22:       Update(LRU_List, W_i)
23:     end if
24:   end if
25: end procedure
```

Algorithm 2 provides a detailed description of the buffer destaging scheme in GCaR. Once a write request arrives, the GCaR scheme will check whether it hits the buffer. If yes, the buffered data will be updated. Otherwise, the buffer free space availability will be checked. If the buffer is full, GCaR will destage data blocks to make free space for the new write data. The data blocks at the LRU tail will be

TABLE 2
The Default SSD Model Parameters

Parameter	Value
Total Capacity	28 GB
Reserved Free Blocks	15%
Minimum Free Blocks	5%
Cleaning Policy	Greedy
Flash Chip Elements	7
Planes Per Package	8
Blocks Per Plane	1024
Pages Per Block	64
Page Size	4 KB
Page Read Latency	25 us
Page Write Latency	200 us
Block Erase Latency	1.5 ms

checked to make sure that the destaged data blocks do not belong to the flash chips that are in the GC state. Once determined, the evicted data blocks will be written to the flash chips that are not in the GC state. The new write data will be kept in the buffer and the LRU list is updated. If the buffer has free space, the new write data will be written to the buffer directly and the LRU list is updated.

4 Performance Evaluations

In this section, we first describe the experimental setup and methodology. Then we evaluate the performance of GCaR through both benchmark-driven and trace-driven evaluations.

4.1 Experimental Setup and Methodology

To evaluate the efficiency of our proposed GCaR scheme, we have implemented a prototype of the GCaR scheme by integrating it into an open-source SSD simulator developed by Microsoft Research (MSR) [1]. The MSR SSD simulator, an extension of Disksim from the Parallel Data Lab of CMU [36], has been released to the public and widely used to evaluate the performance of the SSD-based storage systems in many studies [1], [8], [37]. In this paper, we extend the original Disksim and the MSR SSD simulator to implement our proposed GCaR scheme. The main part of GCaR is implemented in the memory layer, *disksim_cachemem.c*, in MSR SSD-extended DiskSim simulator. Moreover, *ssd.c* is also extended to provide the internal information to the cache layer. We implemented the GCaR prototype by adding 1026 LOC to MSR SSD-extended DiskSim simulator. The values of the SSD specific parameters used in the simulator are shown in Table 2. To have a fair comparison among different schemes, the simulated SSDs are filled up before each experiment.

We use both the synthetic traces and the realistic enterprise-scale workloads to study the performance impact of the different buffer cache schemes. The synthetic workloads allow us to flexibly vary parameters such as request size, inter-arrival time of requests, read access probability, and sequentiality. The default values of the parameters that we use in our experiments are shown in Table 3. The Financial workload is collected from OLTP applications running at a large financial institution [38] and the other six realistic enterprise-scale workloads were collected from the Microsoft Cambridge Research [39]. The main workload characteristics of these traces are summarized in Table 4.

TABLE 3
The Default Parameters of Synthetic Traces

Parameter	Value
Request size	32 KB
Inter-arrival time	4 ms
Probability of sequential access	0
Probability of read access	0.2

TABLE 4
The Realistic Workload Characteristics

Traces	Write Ratio	Total I/Os	Average Request Size
Financial1	76.8%	5,334,981	8.5 KB
Prn_0	89.2%	5,585,886	22.2 KB
Prxy_0	97.1%	5,000,000	6.8 KB
Hm_0	64.5%	3,993,300	9.2 KB
Wdev_0	80.0%	907,930	16.0 KB
Mds_0	88.1%	1,022,042	14.1 KB
Rsrch_0	90.7%	1,292,454	17.3 KB

In the evaluations, we have integrated our proposed GCaR scheme with the LRU, CFLRU [35], and BPLRU [26] schemes, with the resulting GCaR-based schemes being labeled GCaR-LRU, GCaR-CFLRU, and GCaR-BPLRU, respectively. We compare the performance of these GCaR based schemes with LRU, CFLRU, and BPLRU schemes in terms of the average response time. The LRU scheme is the baseline for all the other schemes. The main features of the CFLRU and BPLRU schemes are summarized below.

- *CFLRU* (Clean First LRU) [35] is a buffer cache management algorithm for flash storage. It was proposed to exploit the asymmetric performance of flash memory read and write operations. It attempts to choose a clean page rather than a dirty one as a victim because writing cost of the latter is much more expensive.
- *BPLRU* (Block Padding LRU) [26] aims to improve the random-write performance of SSDs. It combines three key techniques, block-level LRU management, page padding, and LRU compensation to convert random write requests to sequential ones.

4.2 Performance Results

Synthetic Workloads. Fig. 6 shows the average response times, normalized to that of the baseline with a buffer size of 1 MB, of the different schemes as a function of the buffer size, driven by the synthetic workloads. We can see that GCaR-based schemes reduce the average response times by 35.9, 32.1, and 30.6 percent on average compared with the LRU, CFLRU, and BPLRU schemes respectively. It is clear that, by avoiding issuing GC-conflicting user requests to the flash chips in the GC state, all the GCaR-based schemes are able to significantly reduce user response times. The results indicate that GC operations have a significant impact on the user response times. On the other hand, as expected, the average response times decrease as the buffer size increases. When the buffer size is larger than 4 MB, the response times become stable. However, the improvements of the GCaR-based schemes over their original ones are also consistent.

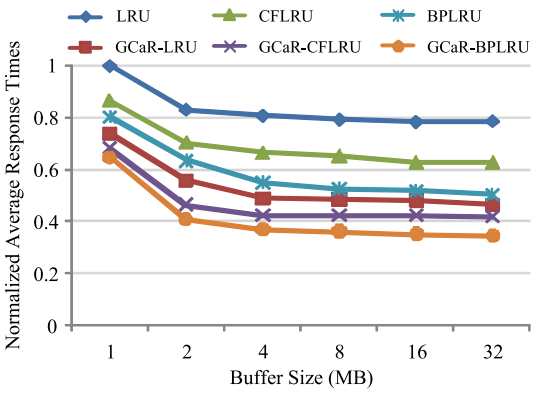

Fig. 6. The average response times, normalized to that of the LRU scheme with a 1MB buffer, under the synthetic workloads.

For example, GCaR-LRU reduces the average response time of the LRU scheme by 40.7 percent when the buffer size is 32 MB. With a larger buffer size, GCaR-LRU scheme can keep data blocks belonging to the flash chip in the GC state in memory without scarifying too much data access locality of other blocks.

To better understand the reasons behind the above results, we examine the cache hit rate and GC count measures collected in the simulation study, as shown in Figs. 7 and 8. Fig. 7 shows that the LRU scheme has the highest cache hit rate under the different buffer sizes, though the differences are not significant. The reason is obvious. Among all the schemes, the LRU scheme is the only one that tries exclusively to exploit the access locality to improve the cache hit rate. The other schemes try to improve the GC efficiency or reduce the miss penalty. Thus, the access locality may be sacrificed to some extend. However, their designs are based on the LRU scheme, the overall hit rates are not reduced significantly. On the other hand, the cache hit rates are not consistent with the average access times, as shown in Fig. 6. This phenomenon further implies that cache hit rate is not the only influential factor affecting the overall performance. It is also the reason why cache hit rate should not be the only design objective of an effective cache scheme for flash storage.

Fig. 8 shows the GC count of the different cache schemes as a function of the buffer size, driven by the synthetic workloads. First, compared with the LRU scheme, both the CFLRU and BPLRU schemes reduce the GC count significantly under the different buffer sizes, especially when the buffer size is small. BPLRU reduces the total GC count of the LRU scheme by up to 11.2 percent. It confirms that reducing GC count can significantly improve the overall system performance. In other words, the GC operations have a significant impact on the overall system performance. Second, it is interesting to notice that, while the GCaR-based schemes have similar GC counts to their non-GCaR counterparts, the former's average response times are much better than the latter's. The reason is that the GCaR-based schemes do not try to further improve the GC efficiency, but instead try to reduce the interference between the user I/O requests and GC-induced I/O requests.

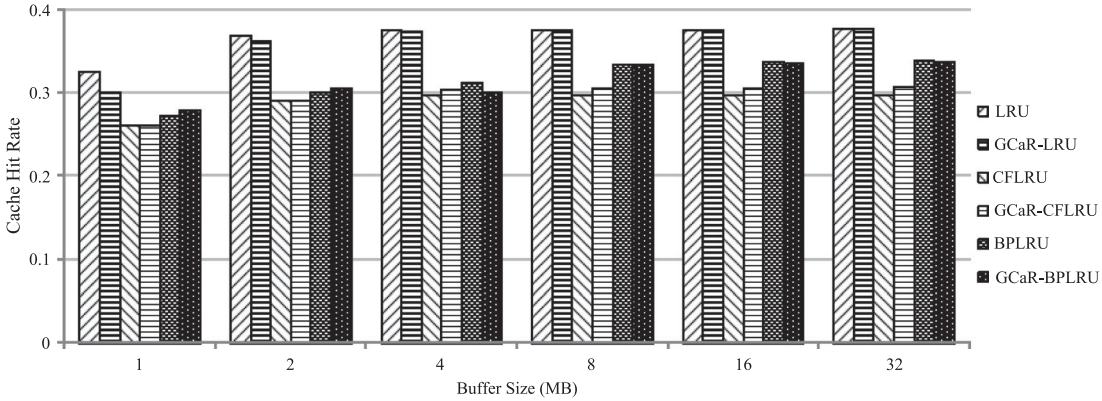

Fig. 7. The cache hit rate of different cache schemes as a function of the buffer size, driven by the synthetic workloads.

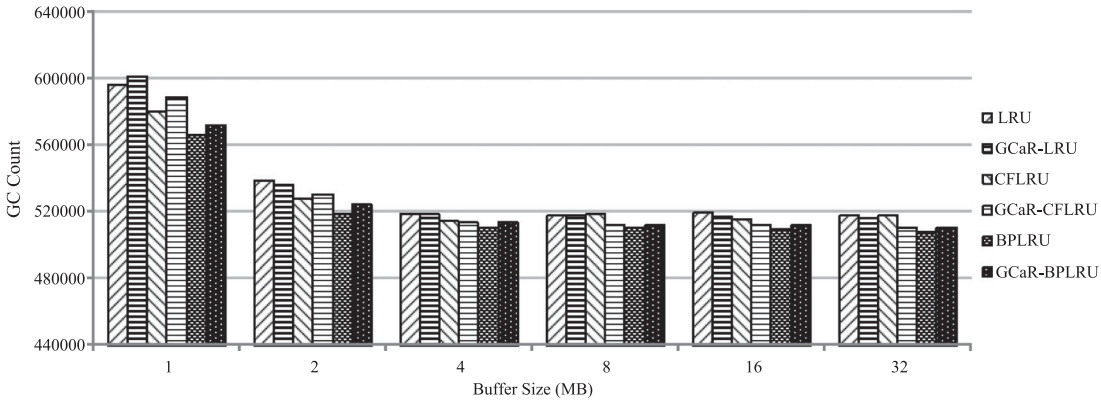

Fig. 8. The GC count of different cache schemes as a function of the buffer size, driven by the synthetic workloads.

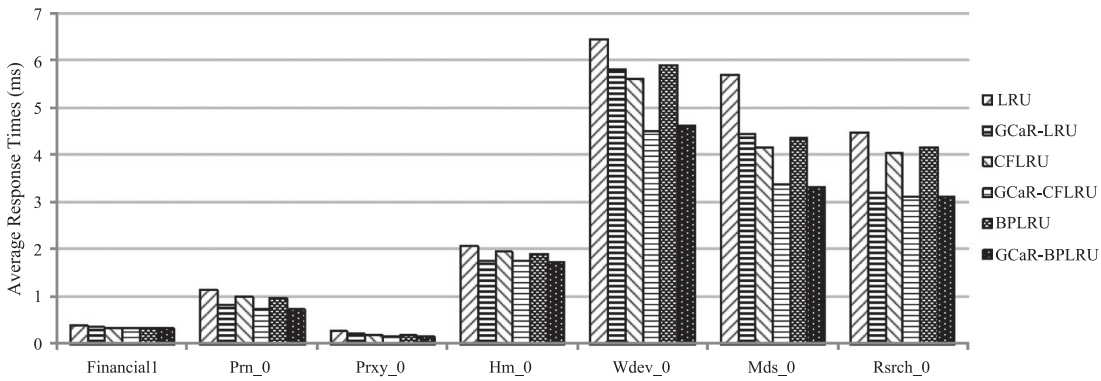

Fig. 9. The average response times under the realistic trace-driven evaluations.

Although the GC efficiency is not further improved by the GCaR-based schemes, the responsiveness to the user I/O requests has been significantly improved. The overall system performance in terms of the average response time is also improved accordingly. Third, the GCaR-based schemes achieve their superiority in overall performance without sacrificing advantages of their non-GCaR counterparts, as shown in Fig. 6. In other words, the GCaR scheme is orthogonal and complementary to these existing schemes. This is because the GCaR scheme only takes the GC activities into the buffer cache design, without changing the original cache algorithm in the normal operational (non-GC) state. The extra resource overhead is minimal. Thus the other schemes can be easily extended to be the GCaR enhanced schemes to improve the system performance.

Real Application Traces. Fig. 9 shows the average response times of the different schemes driven by the seven traces when the buffer size is 1 MB. The GCaR-based cache schemes perform much better than the their non-GCaR counterparts in terms of the average response times. Compared with the LRU scheme, GCaR-LRU reduces the average response times by 8.5, 37.1, 21.9, 17.5, 11.2, 27.5, and 39.5 percent for the Financial1, Prn_0, Prxy_0, Hm_0, Wdev_0, Mds_0, and Rsrch_0 workloads respectively, with an average of 23.3 percent. We also see that both CFLRU and BPLRU perform better than LRU. It further validates that the traditional cache management schemes are not suitable for flash-based storage systems. The reason is that the internal operations and device characteristics of flash memory are different from those of HDDs. For example, the read and write access times are different for a page while the extra erase operation is performed in the unit of a block for flash memory. Moreover, the execution time of an GC operation is one or two orders of magnitude more than that of a write or read operation, which is elaborated in Section 2.1.

On the other hand, we can also see that compared with CFLRU and BPLRU, GCaR-CFLRU and GCaR-BPLRU reduce the average response times by an average of 21.2 and 16.0 percent respectively. Especially for the Prn_0, Prxy_0, Wdev_0, Mds_0, and Rsrch_0 workloads, the average response times of GCaR-CFLRU and GCaR-BPLRU are reduced up to by 34.3 and 25.0 percent respectively. The reason is that for these workloads, the write ratio is high and the request size is large, which means that much more data is written to the flash chips than workloads with low write ratios and small request sizes. Also note that the GC operations driven by these workloads are much more frequent than the other workloads. As a result, the GCaR-based schemes, capable of issuing much fewer GC-conflicting user requests to the chips in the GC state, are able to significantly reduce the overall user response times.

We now examine the cache hit rates of the different schemes driven by the seven traces, normalized to that of the baseline with the LRU scheme. Fig. 10 shows that, similar to the case of synthetic workloads, the LRU scheme outperforms all the schemes in the cache hit rate measure, although its advantage is only marginal by less than 5 percent. The reason is that the other schemes are all build on top of the LRU scheme with some specific optimizations. Thus, the workload locality is captured. On the other hand, compared with the results of the average response times in Fig. 9, the cache hit rate is not consistent with the average

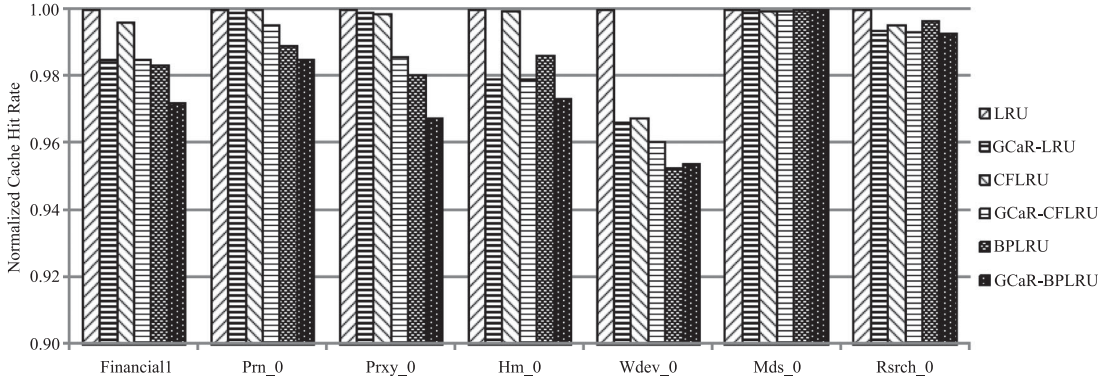

Fig. 10. The cache hit rates, normalized to that of the baseline scheme (LRU), under the trace-driven evaluations.

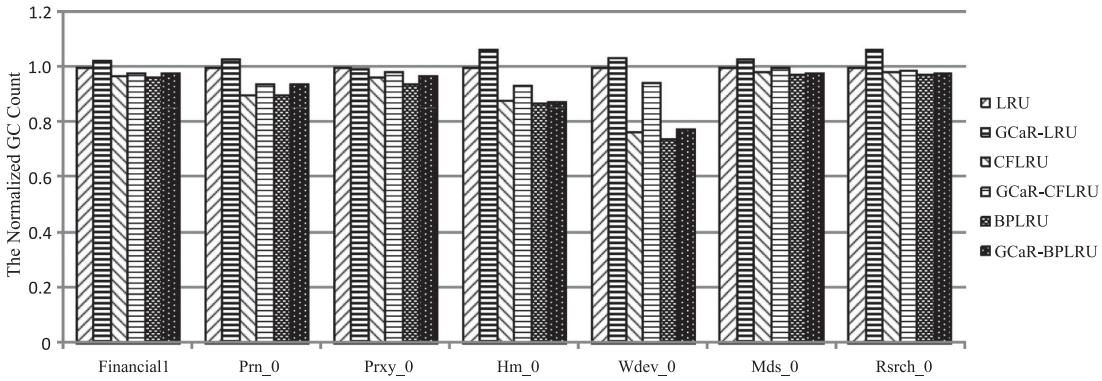

Fig. 11. The GC counts, normalized to that of the baseline scheme (LRU), under the trace-driven evaluations.

response times. The reason is obvious in that the average response times are depended on three factors: the hit time, the miss rate and the miss penalty. The former two factors are almost the same for all the schemes. The miss penalty is not consistent for flash chips, e.g., one or two orders of magnitude higher penalty on flash chips in the GC state than on the non-GC flash chips. Although the difference of hit ratios between GCaR and its non-GCaR counterparts is only less than 2 percent, the response time is significantly different due to the large gap in miss penalties between them. As shown in Equation 1, the average memory access time is not only determined by the hit ratio, but also by the miss penalty, when the hit time is fixed.

Besides the cache hit rates, we also collect the GC counts during the trace-driven evaluations, normalized to that of the baseline with the LRU scheme. Fig. 11 shows that request access delays caused by the GC operations are an influential part of the average response times. Thus, with the real application workloads we obtain consistent results with and the same conclusion as with the synthetic workloads. We can see that compared with the LRU scheme, the CFLRU and BPLRU schemes reduce the GC counts by an average of 8.1 and 9.3 percent respectively, which further confirms that GC operations have a significant impact on the overall system performance. Thus, reducing the GC counts can significantly reduce the average response times. Moreover, we can also see that the GCaR-based schemes have similar GC counts to their non-GCaR counterparts. The reason is that the GCaR-based schemes are based on top of their non-GCaR counterparts with optimizations during the GC active periods. The objective of GCaR-based schemes is to reduce the interference between the user I/O requests and GC-induced I/O requests, but do not try to improve the GC efficiency. It further validates the portability characteristics of the GCaR scheme which is orthogonal and complementary to these existing schemes. The GCaR scheme can be embedded into the LRU, CFLRU and BPLRU schemes to show its efficiency without changing the original cache algorithm in the normal operational (non-GC) state.

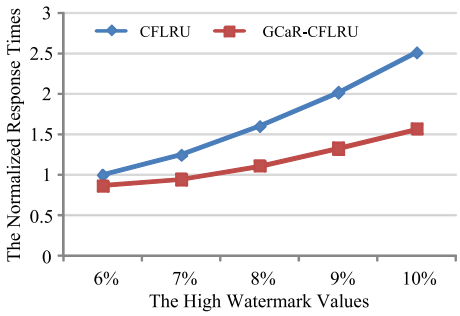

Fig. 12. The average response time, normalized to that of the baseline with CFLRU scheme under 6 percent high watermark, as a function of the high watermark value, driven by the synthetic workloads.

4.3 Sensitivity Study

The cache performance in the presence of GC operations is affected by various factors, such as the high/low watermark (defined below) and the request size. In order to evaluate their impact on the system performance, we carry out several sensitivity studies driven by the synthetic workloads. Here we only report results that compare GCaR-CFLRU with CFLRU for simplicity, since all the other schemes share the same trend.

High Watermark. The high watermark is the threshold on the number of available free blocks in a flash chip, below which a GC operation within chip will be triggered. It is set by default at 5 percent of the total number of blocks in a chip, as shown in Table 2. Fig. 12 shows the average response time, normalized to that of the baseline with CFLRU scheme under 6 percent high watermark, as a function of the high watermark values driven by the synthetic workloads. It shows that the overall response time increases with high watermark value. The reason is that with a higher watermark, the GC operations will be triggered more frequently and thus more blocks will be erased to free up the invalid blocks, which in turn increases the interference between the user IO requests and GC-induced IO requests. However, the GCaR scheme is shown to be much less affected by the change in the high watermark value than the CFLRU scheme. The reason is that the GCaR scheme tries to alleviate the effect of GC operations on the user performance. Thus the GC operation overhead on the system performance in the GCaR-based scheme is much less than the CFLRU scheme.

Request Size. Fig. 13 shows the average response time, normalized to that of the baseline with the CFLRU scheme with a request size of 4 KB, as a function of the request size, driven by the synthetic workloads. It shows that as request size increases, so does the response time. The reasons are two-fold. First, with a larger request size, the transfer time is also longer since the response time is linearly increased

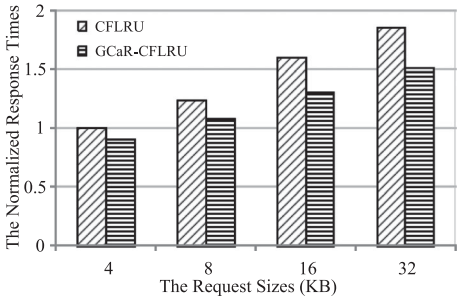

Fig. 13. The average response time, normalized to that of the baseline with CFLRU scheme with a request size of 4KB, as a function of the request size, driven by the synthetic workloads.

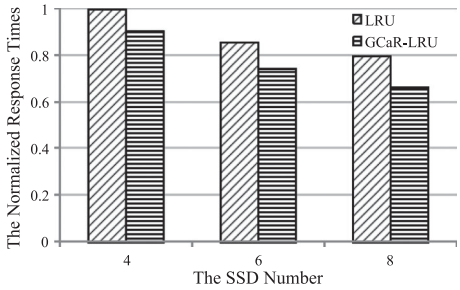

Fig. 14. The average response time, normalized to that of the baseline with the LRU scheme with 4 SSDs and a buffer size of 1 GB, as a function of the number of SSDs, driven by the synthetic workloads.

with the increased request size [40]. Second, with a larger request size, the overall written data size is increased. Thus the available free space is also decreased much faster, which causes the GC operations to be triggered more frequently. Nevertheless, the GCaR-CFLRU scheme consistently outperforms the CFLRU scheme in overall response time regardless of the request size.

4.4 Extendibility Study

The GC operations within SSDs have a significant impact on the SSD performance, thus leading to the performance variability in the RAIS system. The access times to the GC SSDs and the non-GC SSDs are different. The buffer cache management for the RAIS should also be aware of the underline GC activities to address the performance variability. To examine the extendibility of the GCaR scheme on the RAIS system, we also conduct the evaluations on a RAIS5 system consisting of multiple SSDs, each with a 64 GB by using the Disksim simulator.

Fig. 14 shows the average response time, normalized to that of the baseline with the LRU scheme with 4 SSDs and a buffer size of 1 GB, as a function of the number of SSDs, driven by the synthetic workloads. We can see that GCaR-LRU reduces the average response times by up to 19.7 percent under 8 SSDs, and with an average of 15.4 percent, compared with the LRU scheme. The reason is that with more SSDs within the RAIS5 system driven by the synthetic workloads, more SSDs will be in the GC states. The performance variability problem will be much more serious due to the much more frequent GC activities within the RAIS5 system. GCaR-LRU is aware of the underline GC activities by keeping the data blocks belong to the GC SSDs in memory which can significantly alleviate the performance degradation problem caused by the interference between the user I/O requests and the GC-induced I/O requests. On the other hand, the results also validate that GCaR can be extended in the other environments to boost the system performance.

5 Related Work

Buffer cache is one of the most important and effective performance optimizations in storage systems [41]. Its management is well studied in the literature and a large number of cache algorithms have been proposed. Most of them aim to improve the HDD-based storage systems. However, due to the different performance characteristics, the cache schemes for HDDs are not suitable for flash-based SSDs [26], [27], [35], [42].

Flash memory has the asymmetric performance for the read and the write operations [1]. CFLRU [35] divides the LRU list into the working region and the clean-first region, and adopts a policy that evicts clean pages preferentially in the clean-first region until the number of pages hit in the working region reaches a suitable level. Their experiments show that CFLRU is able to reduce the average replacement cost by 26 percent in the buffer cache compared to the traditional LRU algorithm. CFLRU reduces the number of writes by trading off the number of reads. However, this is irrelevant when only write requests are involved. The Flash Aware Buffer policy (FAB) [43] groups the data pages belonging to the same erasable block together and manages with an LRU list. A group is moved to the beginning of the list when a buffer in the group is read or updated, or a new buffer is added to the group. When replacing a cache line out of the memory, a group that has the largest number of data blocks is selected as victim. All the dirty data blocks in the victim group are flushed, and all the clean data blocks in it are discarded. The main use of FAB is in portable media player applications in which the majority of write requests are sequential. However, both CFLRU and FAB are not useful for enhancing random write performance.

To address the random-write performance issue, BPLRU [26] manages a LRU list in the unit of NAND-flash erasable block and pads a log block with some clean pages from the data block to reduce the number of full merges. However, it fails to consider the cost of extra padding operations, which can degrade the performance significantly and become a performance bottleneck particularly when the buffer is much smaller than the working set. PUDLRU (Predicted average Update Distance LRU) [44] partly solves this problem by carefully considering the average update distance and the fullness of the buffered blocks and then selecting one as victim. The extended GC-ARM [27] dynamically destages either contiguous pages in a block as a whole or a single page from the write buffer based on the benefit to improve GC efficiency.

Some studies also try to dynamically allocate buffer space between mapping table and user data to improve system performance. Hyotaek et al. [45] propose an adaptive RAM partitioning scheme for SSDs, which adaptively tunes the space ratio of the data buffer and mapping table according to the workload characteristics. Hu et al. propose PASS [46], which judiciously and actively writes an appropriate amount of the buffered data back to flash by exploiting the light-traffic periods of workloads and the parallelism inside SSDs to

TABLE 5
The Related Studies with GCaR and Their Differences

Schemes	Devices	Objective	Comments
PB-LRU [31]	HDDs	Power efficiency	Miss penalties on the active disks and sleeping disks are different
VDF [32]	HDDs	Availability	Miss penalties on the active disks and faulty disks are different
BPLRU [26]	SSDs	Performance	Improving GC efficiency proactively but unaware of run-time GCs
GC-ARM [27]	SSDs	Performance & Reliability	Improving GC efficiency proactively but unaware of run-time GCs
GCaR	SSDs	Performance	Miss penalties in run-time GC-active flash chips and GC-inactive flash chips are different

avoid real-time flash operations and buffer overflow during bursty-traffic periods. These space management schemes only adaptively change the space allocation ratio between the mapping table and the write data according the workload characteristics. Thus the cache management algorithms for the user data are not affected.

All these schemes work well in the normal state when there are no GC operations in process within the flash chips. Whenever a GC operation is triggered on a particular flash chip, most of the assumptions made for these schemes are no longer valid. For example, the response time of a read request on an ongoing GC-based flash chip is much longer than that of a write request on a flash chip without GC operations. Thus, the CFLRU scheme, for example, will lose its key efficiency. Other optimizations, such as BPLRU, PUDLRU, and GC-ARM, only proactively improve the GC efficiency by reducing the amount of random write data blocks. Their main and shared objective is to improve the GC efficiency by optimizing the GC workflow. However, GCaR improves the GC efficiency by avoiding the contention between user requests and GC-induced requests. Dynamic allocation scheme [47] is trying to avoid the resource contention between the user I/O requests and the GC-induced requests. However, dynamic allocation needs extra memory to store the mapping information and changes the data layout on flash chips, which may affect the GC efficiency. Moreover, our GCaR scheme works not only for write requests, but also for read requests. Thus, our proposed GCaR scheme is orthogonal to and can be easily incorporated into the most existing cache management algorithms to further improve system performance.

Table 5 presents a summary comparison among the related studies and GCaR. Since the main objective of most cache algorithms proposed for HDD-based storage systems is performance, once the objective is changed, their management is also changed. For example, PB-LRU [31] takes the power saving rather performance as the design objective and dynamically allocates different buffer space for different disks based on the power state, because the miss penalties to the active disk and sleeping disk are different. VDF and Shaper [32], [33] are cache management schemes for disk arrays under disk failure. In this condition, the miss penalties to the active disks and the faulty disk are different. Thus they give higher priority for the data blocks belonging to the faulty disk in buffer cache to alleviate the disk contention between the user I/O requests and the reconstruction-induced I/O requests. Our proposed GCaR is inspired by the PB-LRU, VDF and Shaper schemes, but works for the flash-based SSDs with different design objectives.

6 Conclusion

With the widening gap between the speeds of the processor/memory and the HDDs, the I/O access latency has become the system performance wall. Flash-based SSDs have emerged to be a promising technology to bridge this gap to reduce the access latency in the high performance computing environment and enterprise data centers. However, GC is a significant performance concern for flash-based storage systems in enterprise and HPC environments. In this paper, we propose a Garbage Collection aware Replacement policy, called GCaR, to improve the performance of flash-based SSDs. The basic idea is to give higher priority to caching the data blocks belonging to the flash chips that are in the GC state. This substantially lessens the contentions between the user I/O operations and the GC-induced I/O operations. We have integrated GCaR into the SSD extended Disksim simulator and conducted extensive evaluations. The performance results show that GCaR can significantly improve the storage performance by substantially reducing the aforementioned contentions.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61472336 and No. 61402385, US NSF under Grant No. NSF-CNS-1116606 and NSF-CNS-1016609. This work is also sponsored by Huawei Innovation Research Program. Dr. Mao is the corresponding author. This is an extended version of our manuscript published in the Proceedings of the 30th International Conference on Supercomputing (ICS'16), Istanbul, Turkey, Jun. 1-3, 2016.

References

[1] N. Agrawal, V. Prabhakaran, T. Wobber, J. Davis, M. Manasse, and R. Panigrahy, "Design tradeoffs for SSD performance," in *Proc. USENIX Annu. Technical Conf.*, Jun. 2008, pp. 57–70.

[2] F. Chen, D. A. Koufaty, and X. Zhang, "Hystor: Making the best use of solid state drives in high performance storage systems," in *Proc. Int. Conf. Supercomputing*, Jun. 2011, pp. 22–32.

[3] J. Hu, H. Jiang, and P. Manden, "Understanding performance anomalies of SSDs and their impact in enterprise application environment," in *Proc. 12th ACM SIGMETRICS/PERFORMANCE Joint Int. Conf. Meas. Model. Comput. Syst.*, Jun. 2012, pp. 415–416.

[4] M. Jung, W. Choi, J. Shalf, and M. Kandemir, "Triple-A: A non-SSD based autonomic all-flash array for scalable high performance computing storage systems," in *Proc. 19th Int. Conf. Architectural Support Program. Languages Operating Syst.*, Mar. 2014, pp. 441–454.

[5] B. Mao, et al., "HPDA: A hybrid parity-based disk array for enhanced performance and reliability," *ACM Trans. Storage*, vol. 8, no. 1, pp. 1–20, 2012.

[6] W. Wang, T. Xie, and D. Zhou, "Understanding the impact of threshold voltage on MLC flash memory performance and reliability," in *Proc. 28th ACM Int. Conf. Supercomputing*, Jun. 2014, pp. 201–210.

[7] L. Grupp, J. Davis, and S. Swanson, "The bleak future of NAND flash memory," in *Proc. 10th USENIX Conf. File Storage Technol.*, Feb. 2012, pp. 17–24.

[8] J. Lee, Y. Kim, G. Shipman, S. Oral, F. Wang, and J. Kim, "A semi-preemptive garbage collector for solid state drives," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw.*, Apr. 2011, pp. 12–21.

[9] A. M. Caulfield, et al., "Understanding the impact of emerging non-volatile memories on high-performance, IO-intensive computing," in *Proc. ACM/IEEE Int. Conf. High Perform. Comput. Netw., Storage Anal.*, Nov. 2010, pp. 1–11.

[10] A. Gupta, Y. Kim, and B. Urgaonkar, "DFTL: A flash translation layer employing demand-based selective caching of page-level address Mappings," in *Proc. 14th Int. Conf. Architectural Support Program. Languages Operating Syst.*, Mar. 2009, pp. 229–240.

[11] D. Narayanan, E. Thereska, A. Donnelly, S. Elnikety, and A. Rowstron, "Migrating server storage to SSDs: Analysis of tradeoffs," in *Proc. 4th ACM Eur. Conf. Comput. Syst.*, Mar. 2009, pp. 145–158.

[12] S. Oral, F. Wang, D. A. Dillow, G. M. Shipman, and R. Miller, "Efficient object storage journaling in a distributed parallel file system," in *Proc. 8th USENIX Conf. File Storage Technol.*, Feb. 2010, pp. 143–154.

[13] Y. Kim, R. Gunasekaran, G. M. Shipman, D. A. Dillow, Z. Zhang, and B. W. Settlemyer, "Workload characterization of a leadership class storage," in *Proc. 5th Petascale Data Storage Workshop*, Nov. 2010, pp. 1–5.

[14] Y. Kim, S. Oral, G. M. Shipman, J. Lee, D. A. Dillow, and F. Wang, "Harmonia: A globally coordinated garbage collector for arrays of solid-state drives," in *Proc. IEEE 27th Symp. Mass Storage Syst. Technol.*, May 2011, pp. 1–12.

[15] M. Jung, R. Prabhakar, and M. Kandemir, "Taking garbage collection overheads off the critical path in SSDs," in *Proc. 13th Int. Middleware Conf.*, Dec. 2012, pp. 164–186.

[16] F. Chen, T. Luo, and X. Zhang, "CAFTL: A content-aware flash translation layer enhancing the lifespan of flash memory based solid state drives," in *Proc. 9th USENIX Conf. File Stroage Technol.*, Feb. 2011, pp. 77–90.

[17] A. Gupta, R. Pisolkar, B. Urgaonkar, and A. Sivasubramaniam, "Leveraging value locality in optimizing NAND flash-based SSDs," in *Proc. 9th USENIX Conf. File Stroage Technol.*, Feb. 2011, pp. 91–103.

[18] G. Wu and X. He, "Delta FTL: Improving SSD lifetime via exploiting content locality," in *Proc. 7th ACM Eur. Conf. Comput. Syst.*, Apr. 2012, pp. 253–266.

[19] Z. Chen and K. Shen, "OrderMergeDedup: Efficient, failure-consistent deduplication on flash," in *Proc. 14th Usenix Conf. File Storage Technol.*, Feb. 2016, pp. 291–299.

[20] X. Zhang, J. Li, H. Wang, K. Zhao, and T. Zhang, "Reducing solid-state storage device write stress through opportunistic in-place delta compression," in *Proc. 14th Usenix Conf. File Storage Technol.*, Feb. 2016, pp. 111–124.

[21] Y. Hu, H. Jiang, D. Feng, L. Tian, H. Luo, and S. Zhang, "Performance impact and interplay of SSD parallelism through advanced commands, allocation strategy and data granularity," in *Proc. Int. Conf. Supercomputing*, Jun. 2011, pp. 96–107.

[22] M. Jung and M. Kandemir, "Sprinkler: Maximizing resource utilization in many-chip solid state disks," in *Proc. IEEE 20th Int. Symp. High Perform. Comput. Archit.*, Feb. 2014, pp. 524–535.

[23] B. Mao, H. Jiang, S. Wu, Y. Fu, and L. Tian, "SAR: SSD assisted restore optimization for deduplication-based storage systems in the cloud," in *Proc. IEEE 7th Int. Conf. Netw., Archit. Storage*, Jun. 2012, pp. 328–337.

[24] M. Huang and L. Men, "Improving the performance of on-board cache for flash-based solid-state drives," in *Proc. IEEE 7th Int. Conf. Netw. Archit. Storage*, Aug. 2012, pp. 283–287.

[25] X. Hu, et al., "LAMA: Optimized locality-aware memory allocation for key-value cache," in *Proc. USENIX Annu. Technical Conf.*, Jun. 2015, pp. 57–69.

[26] H. Kim and S. Ahn, "BPLRU: A buffer management scheme for improving random writes in flash storage," in *Proc. 6th USENIX Conf. File Storage Technol.*, Feb. 2008, pp. 239–252.

[27] J. Hu, H. Jiang, L. Tian, and L. Xu, "GC-ARM: Garbage collection-aware RAM management for flash based solid state drives," in *Proc. IEEE 7th Int. Conf. Netw. Archit. Storage*, Jun. 2012, pp. 134–143.

[28] M. Jung, W. Choi, S. Srikantaiah, J. Yoo, and M. Kandemir, "HIOS: A host interface I/O scheduler for solid state disks," in *Proc. ACM/IEEE 41st Int. Symp. Comput. Archit.*, Jun. 2014, pp. 289–300.

[29] S. Yan, et al., "Tiny-tail flash: Near-perfect elimination of garbage collection tail latencies in NAND SSDs," in *Proc. USENIX Conf. File Storage Technol.*, Feb. 2017, pp. 15–28.

[30] Understanding SSD performance project of SNIA. [Online]. Available: http://www.snia.org/forums/sssi/pts

[31] Q. Zhu, A. Shankar, and Y. Zhou, "PB-LRU: A self-tuning power aware storage cache replacement algorithm for conserving disk energy," in *Proc. 18th Annu. Int. Conf. Supercomputing*, Jun. 2004, pp. 79–88.

[32] S. Wan, et al., "Victim disk first: An asymmetric cache to boost the performance of disk arrays under faulty conditions," in *Proc. USENIX Annu. Technical Conf.*, Jun. 2011, pp. 173–185.

[33] S. Wu, B. Mao, D. Feng, and J. Chen, "Availability-aware cache management with improved RAID reconstruction performance," in *Proc. 13th IEEE Int. Conf. Comput. Sci. Eng.*, Dec. 2010, pp. 229–236.

[34] Y. Kim, J. Lee, S. Oral, D. Dillow, F. Wang, and G. Shipman, "Coordinating garbage collectionfor arrays of solid-state drives," *IEEE Trans. Comput.*, vol. 63, no. 4, pp. 888–901, Apr. 2014.

[35] S. Park, D. Jung, J. Kang, J. Kim, and J. Lee, "CFLRU: A replacement algorithm for flash memory," in *Proc. Int. Conf. Compilers Archit. Synthesis Embedded Syst.*, Oct. 2006, pp. 234–241.

[36] J. Bucy, J. Schindler, S. Schlosser, and G. Ganger, "The DiskSim simulation environment version 4.0 reference manual," Carnegie Mellon University, Tech. Rep. CMU-PDL-08-101, May 2008.

[37] G. Wu and B. He, "Reducing SSD read latency via NAND flash program and erase suspension," in *Proc. 10th USENIX Conf. File Storage Technol.*, Feb. 2012, pp. 117–123.

[38] UMass trace repository. [Online]. Available: http://traces.cs.umass.edu/index.php/Storage/Storage

[39] Block I/O traces in SNIA. [Online]. Available: http://iotta.snia.org/tracetypes/3

[40] B. Mao and S. Wu, "Exploiting request characteristics and internal parallelism to improve SSD performance," in *Proc. 33rd IEEE Int. Conf. Comput. Des.*, Oct. 2015, pp. 447–450.

[41] W. W. S. Hsu and A. J. Smith, "The performance effect of I/O optimizations and disk improvements," *IBM J. Res. Develop.*, vol. 48, no. 2, pp. 255–289, Mar. 2004.

[42] S. Wu, Y. Lin, B. Mao, and H. Jiang, "GCaR: Garbage collection aware cache management with improved performance for flash-based SSDs," in *Proc. Int. Conf. Supercomputing*, Jun. 2016, pp. 1–12.

[43] H. Jo, J. Kang, S. Park, J. Kim, and J. Lee, "FAB: Flash-aware buffer management policy for portable media players," *IEEE Trans. Consumer Electron.*, vol. 52, no. 2, pp. 485–493, May 2006.

[44] J. Hu, H. Jiang, L. Tian, and L. Xu, "PUD-LRU: An erase-efficient write buffer management algorithm for flash memory SSD," in *Proc. IEEE Int. Symp. Model. Anal. Simulation Comput. Telecommun. Syst.*, Aug. 2010, pp. 69–78.

[45] H. Shim, B. Seo, J. Kim, and S. Maeng, "An adaptive partitioning scheme for DRAM-based cache in solid state drives," in *Proc. IEEE 26th Symp. Mass Storage Syst. Technol.*, May 2010, pp. 1–12.

[46] Y. Hu, H. Jiang, D. Feng, H. Luo, and L. Tian, "PASS: A proactive and adaptive SSD buffer scheme for data-intensive workloads," in *Proc. IEEE Int. Conf. Netw. Archit. Storage*, Aug. 2015, pp. 54–63.

[47] A. Tavakkol, M. Arjomand, and H. Sarbazi-Azad , "Unleashing the potentials of dynamism for page allocation strategies in SSDs," in *Proc. ACM Int. Conf. Meas. Model. Comput. Syst.*, Jun. 2014, pp. 551–552.

Suzhen Wu received the BSc and PhD degrees in computer science and technology and computer architecture from Huazhong University of Science and Technology, in 2005 and 2010, respectively. She is an associate professor in the Computer Science Department, Xiamen University since August 2014. Her research interests include computer architecture and storage system. She has more than 40 publications in journal and international conferences. Her research has been supported by NSFC, Huawei, Intel, and Inspur. She is a member of the IEEE and the ACM.

Bo Mao received the BSc degree in computer science and technology from Northeast University, in 2005 and the PhD degree in computer architecture from the Huazhong University of Science and Technology, in 2010. His research interests include storage system, flash-based SSDs and disk arrays, data deduplication, cloud storage, and storage reliability. He was a postdoc researcher with the University of Nebraska-Lincoln between 2010 and 2013. After that, he joined in the Software School, Xiamen University and becomes an associate professor since August 2015. He has more than 40 publications in international journals and conferences. His research has been supported by NSFC and Huawei. He is a member of the IEEE, the ACM, and the USENIX.

Yanping Lin received the BSc degree from the Computer Science Department, Xiamen University, in 2014. She is currently working toward the master's degree at Xiamen University. Her research interests mainly lie in storage systems, particularly the flash-based SSDs, and buffer cache management.

Hong Jiang received the BSc degree in computer engineering from Huazhong University of Science and Technology, China, in 1982, the MASc degree in computer engineering from the University of Toronto, Canada, in 1987, and the PhD degree in computer science from Texas A&M University, in 1991. He is currently chair and Wendell H. Nedderman Endowed professor of Computer Science and Engineering Department, University of Texas at Arlington. Prior to joining UTA, he served as a program director at National Science Foundation (2013.1-2015.8) and he was with the University of Nebraska-Lincoln since 1991, where he was Willa Cather professor of computer science and engineering. His present research interests include computer architecture, computer storage systems and parallel I/O, high performance computing, big data computing, cloud computing, and performance evaluation. He has more than 200 publications in major journals and international Conferences in these areas, and his research has been supported by NSF, DOD, the State of Texas, and the State of Nebraska. He is a fellow of the IEEE, the Member of ACM, and the USENIX.