Characterizing the performance of Flash memory storage devices and its impact on algorithm design

Deepak Ajwani, Itay Malinger,
Ulrich Meyer, Sivan Toledo

MPI–I–2008–1-001 January 2008

Authors' Addresses

Deepak Ajwani
Max-Planck-Institut für Informatik
Stuhlsatzenhausweg 85
66123 Saarbrücken
Germany

Itay Malinger
Tel Aviv University,
Tel Aviv, Israel

Ulrich Meyer
Johann Wolfgang Goethe Universität,
Frankfurt a.M.,
Germany

Sivan Toledo
Massachusetts Institute of Technology,
Massachusetts,
USA

Abstract

Initially used in digital audio players, digital cameras, mobile phones, and USB memory sticks, flash memory may become the dominant form of end-user storage in mobile computing, either completely replacing the magnetic hard disks or being an additional secondary storage. We study the design of algorithms and data structures that can exploit the flash memory devices better. For this, we characterize the performance of NAND flash based storage devices, including many solid state disks. We show that these devices have better random read performance than hard disks, but much worse random write performance. We also analyze the effect of misalignments, aging and past I/O patterns etc. on the performance obtained on these devices. We show that despite the similarities between flash memory and RAM (fast random reads) and between flash disk and hard disk (both are block based devices), the algorithms designed in the RAM model or the external memory model do not realize the full potential of the flash memory devices. We later give some broad guidelines for designing algorithms which can exploit the comparative advantages of both a flash memory device and a hard disk, when used together.

Keywords

Flash memory, external memory model, experimental algorithmics

Contents

1 Introduction

Flash memory is a form of non-volatile computer memory that can be electrically erased and reprogrammed. Flash memory devices are lighter, more shock resistant, consume less power and hence are particularly suited for mobile computing. Initially used in digital audio players, digital cameras, mobile phones, and USB memory sticks, flash memory may become the dominant form of end-user storage in mobile computing: Some producers of notebook computers have already launched models (Apple MacBook Air, Sony Vaio UX90, Samsung Q1-SSD and Q30-SSD) that completely abandon traditional hard disks in favor of flash memory (also called solid state disks). Market research company In-Stat predicted in July 2006 that 50% of all mobile computers would use flash (instead of hard disks) by 2013.

Frequently, the storage devices (be it hard disks or flash) are not only used to store data but also to actually compute on it if the problem at hand does not completely fit into main memory (RAM); this happens on both very small devices (like PDAs used for online route planning) and high-performance compute servers (for example when dealing with huge graphs like the web). Thus, it is important to understand the characteristics of the underlying storage devices in order to predict the real running time of algorithms, even if these devices are used as an external memory. Traditionally, algorithm designers have been assuming a uniform cost access to any location in the storage devices. Unfortunately, real architectures are becoming more and more sophisticated, and will become even more so with the advent of flash devices. In case of hard disks, the access cost depends on the current position of the disk-head and the location that needs to be read/written. This has been well researched; and there are good computation models such as the external memory model [1] or the cache-oblivious model [5] that can help in realistic analysis of algorithms that run on hard disks. This report attempts to characterize the performance (read/writes; sequential/random) of flash memory devices; to see the effects of random writes, misalignment, and aging etc. on the access cost and its implications on the real running time of basic algorithms.

External memory model.

The external memory model (or the I/O model) proposed by Aggarwal and Vitter [1] is one of the most commonly used model when analyzing the performance of algorithms that do not fit in the main memory and have to use the hard disk. It assumes a single central processing unit and two levels of memory hierarchy. The internal memory is fast, but has a limited size of M words. In addition, we have an external memory which can only be accessed using I/Os that move B contiguous words between internal and external memory. At any particular time stamp, the computation can only use the data already present in the internal memory. The measure of performance of an algorithm is the number of I/Os it performs.

State of the art for flash memories.

Recently, there has been growing interest in using flash memories to improve the performance of computer systems [3, 8, 10]. This trend includes the experimental use of flash memories in database systems [8, 10], in Windows Vista's use of USB flash memories as a cache (a feature called ReadyBoost), in the use of flash memory caches in hard disks (e.g., Seagate's Momentus 5400 PSD hybrid drives, which include 256 MB on the drive's controller), and in proposals to integrate flash memories into motherboards or I/O busses (e.g., Intel's Turbo Memory technology).

Most previous algorithmic work on flash memory concerns *operating system* algorithms and data structures that were designed to efficiently deal with flash memory cells wearing out, e.g., block-mapping techniques and flash-specific file systems. A comprehensive overview on these topics was recently published by Gal and Toledo [6]. The development of application algorithms tuned to flash memory is in its absolute infancy. We are only aware of very few published results beyond file systems and wear leveling:

Wu et al. [11, 12] proposed flash-aware implementations of B-trees and R-trees without file system support by explicitly handling block-mapping within the application data structures.

Goldberg and Werneck [7] considered point-to-point shortest-path computations on pocket PCs where preprocessed input graphs (road networks) are stored on flash-memory; due to space-efficient internal-memory data-structures and locality in the inputs, data manipulation remains restricted to internal memory, thus avoiding difficulties with unstructured flash memory write accesses.

Goals.

Our first goal is to see how standard algorithms and data structures for basic algorithms like scanning, sorting and searching designed in the RAM model or the external memory model perform on flash storage devices. An important question here is whether these algorithms can effectively use the advantages of the flash devices (such as faster random read accesses) or there is a need for a fundamentally different model for realizing the full potential of these devices.

Our next goal is to investigate why these algorithms behave the way they behave by characterizing the performance of more than 20 different low-end and high-end flash devices under typical access patterns presented by basic algorithms. Such a characterization can also be looked upon as a first step towards obtaining a model for designing and analyzing algorithms and data structures that can best exploit flash memory. Previous attempts [8, 10] at characterizing the performance of these devices reported measurements on a small number of devices (1 and 2, respectively), so it is not yet clear whether the observed behavior reflects the flash devices, in general. Also, these papers didn't study if these devices exhibit any second-order effects that may be relevant.

Our next goal is to produce a benchmarking tool that would allow its users to measure and compare the relative performance of flash devices. Such a tool should not only allow users to estimate the performance of a device under a given workload in order to find a device with an appropriate cost-effectiveness for a particular application, but also allow quick measurements of relevant parameters of a device that can affect the performance of algorithms running on it.

These goals may seem easy to achieve, but they are not. These devices employ complex logical-to-physical mapping algorithms and complex mechanisms to decide which blocks to erase. The complexity of these mechanisms and the fact that they are proprietary mean that it is impossible to tell exactly what factors affect the performance of a device. A flash device can be used by an algorithm designer like a hard disk (under the external memory or the cache-oblivious model), but its performance may be far more complex.

It is also possible that the flash memory becomes an additional secondary storage device, rather than replacing the hard disk. Our last, but not least, goal is to find out how one can exploit the comparative advantages of both in the design of application algorithms, when they are used together.

Outline.

The rest of the report is organized as follows. In Chapter 2, we show how the basic algorithms perform on flash memory devices and how appropriate are the standard computation models in predicting these performances. In Chapter 3, we present our experimental methodology, and our benchmarking program, which we use to measure and characterize the performance of many different flash devices. We also show the effect of random writes, misalignment, controllers and aging on the performance of these devices. In Chapter 4, we provide an algorithm design framework for the case when flash devices are used together with a hard disk.

2 Implications of flash devices for algorithm design

In this section, we look at how the RAM model and external memory model algorithms behave when running on flash memory devices. In the process, we try to ascertain whether the analysis of algorithms in either of the two models also carry over to the performance of these algorithms obtained on flash devices.

In order to compare the flash memory with DRAM memory (used as main memory), we ran a basic RAM model list ranking algorithm on two architectures - one with 8 GB RAM memory and the other with 2 GB RAM, but 32 GB flash memory. The list ranking problem is that given a list with individual elements randomly stored on disk, find the distance of each element from the head of the list. The sequential RAM model algorithm consists of just hoping from one element to its next, and thereby keeping track of the distances of node from the head of the list. Here, we do not consider the cost of writing the distance labels of each node.

We stored a 2^{30}-element list of long integers (8 Bytes) in a random order, i.e. the elements were kept in the order of a random permutation generated beforehand. While ranking such a list took minutes in RAM, it took days with flash. This is because even though the random reads are faster on flash disks than the hard disk, they are still much slower than RAM. Thus, we conclude that RAM model is not useful for predicting the performance (or even relative performance) of algorithms running on flash memory devices and that standard RAM model algorithms leave a lot to be desired if they are to be used on flash devices.

As Table 2.1 shows, the performance of basic algorithms when running on hard disks and when running on flash disks can be quite different, particularly when it comes to algorithms involving random read I/Os such as binary search on a sorted array. While such algorithms are extremely slow on hard disks necessitating B-trees and other I/O-efficient data structures, they are

Algorithm	Hard Disk	Flash
Generating a random double and writing it	0.2 μs	0.37 μs
Scanning (per double)	0.3 μs	0.28 μs
External memory Merge-Sort (per double)	1.06 μs	1.5 μs
Random read	11.3 ms	0.56 ms
Binary Search	25.5 ms	3.36 ms

Table 2.1: Runtime of basic algorithms when running on Seagate Barracuda 7200.11 hard disk as compared to 32 GB Hama Solid State Disk

acceptably fast on flash devices. On the other hand, algorithms involving write I/Os such as merge sort (with two read and write passes over the entire data) run much faster on hard disk than on flash.

It seems that the algorithms that run on flash have to achieve a different tradeoff between reads and writes and between sequential and random accesses than hard disks. Since the cost of accesses don't drop or rise proportionally over the entire spectrum, the algorithms running on flash devices need to be qualitatively different from the one on hard disk. In particular, they should be able to tradeoff write I/Os at the cost of extra read I/Os. Standard external memory algorithms that assume same cost for reading and writing fail to take advantage of fast random reads offered by flash devices. Thus, there is a need for a fundamentally different model for realistically predicting the performance of algorithms running on flash devices.

3 Characterization of flash memory devices

In order to see why the standard algorithms behave as mentioned before, we characterize more than 20 flash storage devices. This characterization can also be looked at as a first step towards a model for designing and analyzing algorithms and data structures running on flash memory.

3.1 Flash memory

Large-capacity flash memory devices use NAND flash chips. All NAND flash chips have common characteristics, although different chips differ in performance and in some minor details. The memory space of the chip is partitioned into blocks called *erase blocks*. The only way to change a bit from 0 to 1 is to erase the entire unit containing the bit. Each block is further partitioned into *pages*, which usually store 2048 bytes of data and 64 bytes of meta-data (smaller chips have pages containing only 512+16 bytes). Erase blocks typically contain 32 or 64 pages. Bits are changed from 1 (the erased state) to 0 by *programming* (writing) data onto a page. An erased page can be programmed only a small number of times (one to three) before it must be erased again. Reading data takes tens of microseconds for the first access to a page, plus tens of nanoseconds per byte. Writing a page takes hundreds of microseconds, plus tens of nanoseconds per byte. Erasing a block takes several milliseconds. Finally, erased blocks wear out; each block can sustain only a limited number of erasures. The guaranteed numbers of erasures range from 10,000 to 1,000,000. To extend the life of the chip as much as possible, erasures should therefore be spread out roughly evenly over the entire chip; this is called *wear leveling*.

Because of the inability to overwrite data in a page without first erasing the entire block containing the page, and because erasures should be spread out over the chip, flash memory subsystems map *logical block addresses* (LBA) to physical addresses in complex ways [6]. This allows them to accept new data for a given logical address without necessarily erasing an entire block, and it allows them to avoid early wear even if some logical addresses are written to more often than others. This mapping is usually a non-trivial algorithm that uses complex data structures, some of which are stored in RAM (usually inside the memory device) and some on the flash itself.

The use of a mapping algorithm within LBA flash devices means that their performance characteristics can be worse and more complex than the performance of the raw flash chips. In particular, the state of the on-flash mapping and the volatile state of the mapping algorithm can influence the performance of reads and writes. Also, the small amount of RAM can cause the mapping mechanism to perform more physical I/O operations than would be necessary with more RAM.

3.2 Configuration

The tests were performed on many different machines:

- A 1.5GHz Celeron-M with 512M RAM
- A 3.0GHz Pentium 4 with 2GB OF RAM
- A 2.0Ghz Intel dual core T7200 with 2GB OF RAM
- A 2 x Dual-core 2.6 GHz AMD Opteron with 2.5 GB OF RAM

All of these machines were running a 2.6 Linux kernel.

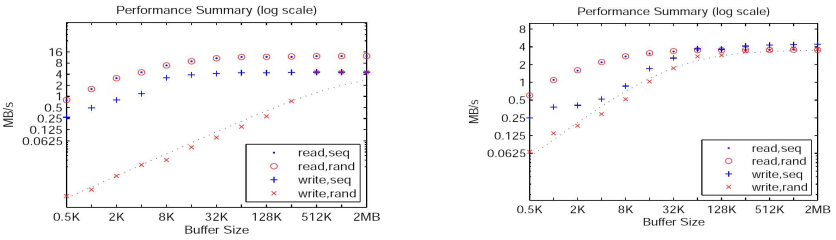
The devices include USB sticks, compact-flash and SD memory cards and solid state disks (of capacities 16GB and 32GB). They include both high-end and low-end devices. The USB sticks were connected via a USB 2.0 interface, memory cards were connected through a USB 2.0 card reader (made by Hama) or PCMCIA interface, and solid state disks with IDE interface were installed in the machines using a 2.5 inch to 3.5 inch IDE adapter and a PATA serial bus.

Our benchmarking tool and methodology.

Standard disk benchmarking tools like `zcav` fail to measure things that are important in flash devices (e.g., write speeds, since they are similar to read speeds on hard disks, or sequential-after-random writes); and commercial benchmarks tend to focus on end-to-end file-system performance, which does not characterize the performance of the flash device in a way that it useful to algorithm designers. Therefore, we decided to implement our own benchmarking program that is specialized (designed mainly for LBA flash devices), but highly flexible and can easily measure the performance of a variety of access patterns, including random and sequential reads and writes, with given block sizes and alignments, and with operation counts or time limits. We provide more details about our benchmarking software and our methodology for measuring devices in Appendix A.

3.3 Result and Analysis

3.3.1 Performance of steady, aligned access patterns.



(a) (b)

Figure 3.1: Performance (in logarithmic scale) of the (a) 1 GB Toshiba Trans-Memory USB flash drive and the (b) 1 GB Kingston compact-flash card.

Figure 3.1 shows the performance of two typical devices under the aligned access patterns. The other devices that we tested varied greatly in the absolute performance that they achieved, but not in the general patterns; all followed the patterns shown in Figures 3.1a and 3.1b.

In all the devices that we tested, small random writes were slower than all the other access patterns. The difference between random writes and other access patterns is particularly large at small buffer sizes, but it is usually still evident even on fairly large block sizes (e.g., 256KB in Figure 3.1a and 128KB in Figure 3.1b). In most devices, small-buffer random writes were at least 10 times slower than sequential writes with the same buffer size, and at least 100 times slower than sequential writes with large buffers. Table 3.1 shows the read/write access time with two different block sizes (512 Bytes

Device		Buffer size 512 Bytes				Buffer size 2 MB			
Name	size	sr	rr	sw	rw	sr	rr	sw	rw
kingston dt secure	512m	0.97	0.97	0.64	0.012	33.14	33.12	14.72	9.85
memorex mini traveldrive	512m	0.79	0.79	0.37	0.002	13.15	13.15	5.0	5.0
toshiba transmemory	512m	0.78	0.78	0.075	0.003	12.69	12.69	4.19	4.14
sandisk u3 cruzer micro	512m	0.55	0.45	0.32	0.013	12.8	12.8	5.2	4.8
m-systems mdrive	1g	0.8	0.8	0.24	0.005	26.4	26.4	15.97	15.97
m-systems mdrive 100	1g	0.78	0.78	0.075	0.002	12.4	12.4	3.7	3.7
toshiba transmemory	1g	0.8	0.8	0.27	0.002	12.38	12.38	4.54	4.54
smi flash device	1g	0.97	0.54	0.65	0.01	13.34	13.28	9.18	7.82
kingston cf card	1g	0.60	0.60	0.25	0.066	3.55	3.55	4.42	3.67
kingston dt elite hs 2.0	2g	0.8	0.8	0.22	0.004	24.9	24.8	12.79	6.2
kingston dt elite hs 2.0	4g	0.8	0.8	0.22	0.003	25.14	25.14	12.79	6.2
memorex td classic 003c	4g	0.79	0.17	0.12	0.002	12.32	12.15	5.15	5.15
120x cf card	8g	0.68	0.44	0.96	0.004	19.7	19.5	18.16	16.15
supertalent solid state flash drive	16g	1.4	0.45	0.82	0.028	12.65	12.60	9.84	9.61
hama solid state disk 2.5” ide	32g	2.9	2.18	4.89	0.012	28.03	28.02	24.5	12.6
ibm deskstar hard drive	60g	5.9	0.03	4.1	0.03	29.2	22.0	24.2	16.2
seagate barracuda 7200.11 hard disk	500g	6.2	0.063	5.1	0.12	87.5	69.6	88.1	71.7

Table 3.1: The tested devices and their performance (in MBps) under sequential and random reads and writes with block size of 512 Bytes and 2 MB.

and 2 MB) for sequential and random accesses on some of the devices that we tested.

We believe that the high cost for random writes of small blocks is because of the LBA mapping algorithm in these devices. These devices partition the virtual and physical address spaces into chunks larger than an erase block; in many cases 512KB. The LBA mapping maps areas of 512KB logical addresses to physical ranges of the same size. On encountering a write request, the system writes the new data into a new physical chunk and keeps on writing contiguously in this physical chunk till it switches to another logical chunk. The logical chunk is now mapped twice. Afterwards, when the writing switches to another logical chunk, the system copies over all the remaining pages in the old chunk and erases it. This way every chunk is mapped once, except for the active chunk, which is mapped twice. On devices that behave like this, the best random-write performance (in seconds) is on blocks of 512KB (or whatever is the chunk size). At that size, the new chunk is written without even reading the old chunk. At smaller sizes, the system still ends up writing 512KB, but it also needs to read stuff from the old location of this chunk, so it is slower. We even found that on some devices, writing randomly 256 or 128KB is slower than writing 512KB, in absolute time.

In most devices, reads were faster than writes in all block sizes. This typical behavior is shown in Figure 3.1a.

Another nearly-universal characteristic of the devices is the fact that sequential reads are not faster than random reads. The read performance does depend on block size, but usually not on whether the access pattern is random or sequential.

The performance in each access pattern usually increases monotonically with the block size, up to a certain saturation point. Reading and writing small blocks is always much slower than the same operation on large blocks.

The exceptions to these general rules are discussed in detail in Appendix B.

Comparison to hard disks.

Quantitatively, the only operation in which LBA flash devices are faster than hard disks is random reads of small buffers. Many of these devices can read a random page in less than a millisecond, sometimes less than 0.5ms. This is at least 10 times faster than current high-end hard disks, whose random-access time is 5-15ms. Even though the random-read performance of LBA flash devices varies, all the devices that we tested exhibited better random-read times than those of hard disks.

In all other aspects, most of the flash devices tested by us are inferior to hard disks. The random-write performance of LBA flash devices is particularly bad and particularly variable. A few devices performed random writes about as fast as hard disks, e.g., 6.2ms and 9.1ms. But many devices were more than 10 times slower, taking more than 100ms per random write, and some took more than 300ms.

Even under ideal access patterns, flash devices we have tested provide smaller I/O bandwidths than hard disks. One flash device reached read throughput approaching 30MB/s and write throughput approaching 25MB/s. Hard disks can achieve well over 100MB/s for both reads and writes. Even disks designed for laptops can achieve throughput approaching 60MB/s. Flash devices would need to improve significantly before they outperform hard disks in this metric. The possible exception to this conclusion is large-capacity flash devices utilizing multiple flash chips, which should be able to achieve high throughput by writing in parallel to multiple chips.

3.3.2 Performance of large number of random writes.

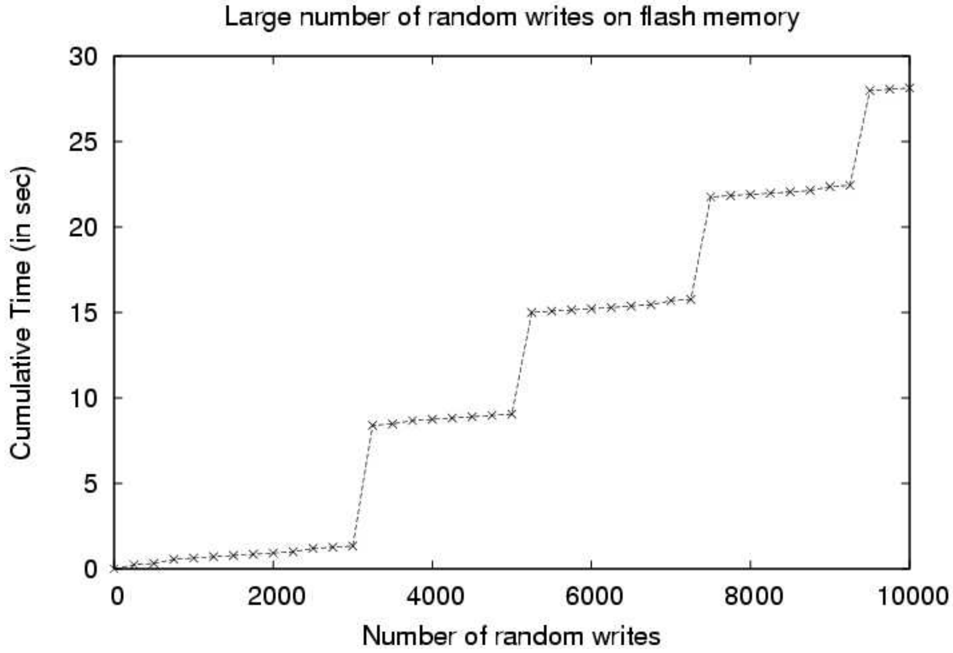


Figure 3.2: Total time taken by large number of random writes on a 32 GB Hama Solid state disk

We observed an interesting phenomenon (Figure 3.2) while performing large number of random writes on a 32 GB Hama (2.5" IDE) solid state disk. After the first 3000 random writes (where one random write is writing a 8-byte real number at a random location in a 8 GB file on flash), we see some spikes in the total running time. Afterwards, these spikes are repeated regularly after every 2000 random writes. This behavior is not restricted to Hama solid state disk and is observed in many other flash devices.

We believe that it is because the random writes make the page table more complex. After a while, the controller rearranges the pages in the blocks to simplify the LBA mapping. This process takes 5-8 seconds while really writing the data on the disk takes less than 0.8 seconds for 2000 random writes, causing the spikes in the total time.

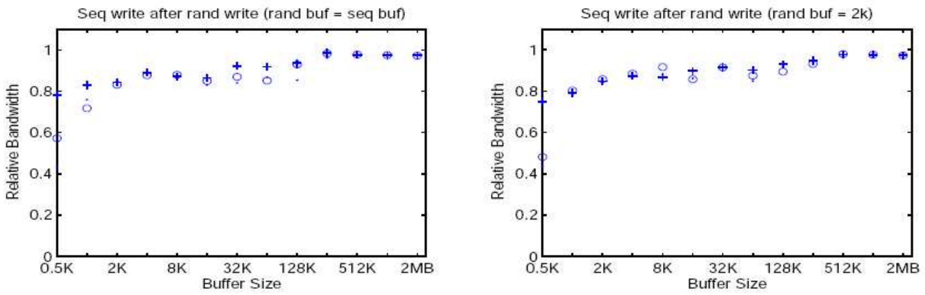


Figure 3.3: Graphs showing the effect of random writes on subsequent sequential writes on Toshiba 1 GB TransMemory USB flash drive.

3.3.3 Effect of random writes on subsequent operations.

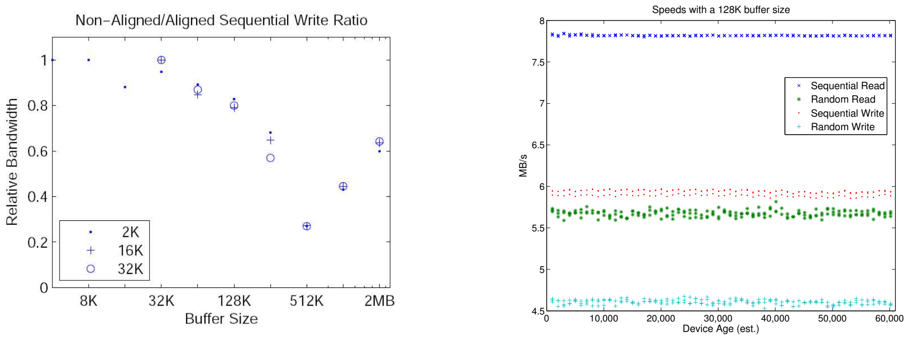
On some devices, a burst of random writes slows down subsequent sequential writes. The effect can last a minute or more, and in rare cases hours (of sustained writing). No such effect was observed on subsequent reads.

Figure 3.3 presents the performance of one such device. In these experiments, we performed t seconds of random writing, for $t = 5, 30$ and 60. We then measured the performance of sequential writes during each 4 second period for the next 120 seconds. The two graphs in Figure 3.3 show the median performance in these 30 4-second periods relative to the steady-state performance of the same pattern (read or write and with the same block size). As we can see, for very small blocks the median performance in the two minutes that follow the random writes can drop by more than a factor of two. Even on larger blocks, performance drops by more than 10%.

More details on the recovery time of flash devices after a random burst of writes (i.e., how long it took the device to recover back to 60% of the median performance in the two minutes following the random writes) are presented in Appendix C.

3.3.4 Effects of misalignment.

On many devices, misaligned random writes achieve much lower performance than aligned writes. In this setting, alignment means that the starting address of the write is a multiple of the block size. We have not observed similar issues with sequential access and with random reads.



(a)

(b)

Figure 3.4: Effect of misalignment on the performance of flash devices

Figure 3.4a shows the ratio between misaligned and aligned random writes. The misalignment is by 2KB, 16KB and 32KB. All of these sizes are at most as large as a single flash page. Many of the devices that we have tested showed some performance drop on misaligned addresses, but the precise effect varied from device to device. For example, the 128MB SuperTalent USB device is affected by misalignment by 2KB but not by misalignments of 16KB or 32KB.

3.3.5 Effects of Aging.

We were not able to detect a significant performance degradation as devices get older (in terms of the number of writes and erasures). Figure 3.4b shows the performance of one device as a function of the number of sequential writes on the entire device. The performance of each access pattern remains essentially constant, even after 60,000 writes. On a different device (512MB Kingston DataTraveler II+), we ran a similar experiment writing more than 320,000 times, exceeding its rated endurance by at least a factor of 3 and did not observe any slowing down with age.

3.3.6 Effect of different controller interfaces.

We connected a compact-flash card via a USB 2.0 interface, PCMCIA interface and an IDE interface (using a card reader) and found that the connecting interface does not affect the relative access patterns (sequential vs. random, read vs. write and the effect of different block sizes) of the flash devices. However, the max read and write bandwidth that we could attain from USB 2.0, PCMCIA and IDE interface are 19.8 MBps (read) with 18.2 MBps (write), 0.95 MBps (read) with 0.95 MBps (write), and 2.16 MBps (read) with 4.38 MBps (write), respectively.

4 Designing algorithms to exploit flash when used together with a hard disk

Till now, we discussed the characteristics of the flash memory devices and the performance of algorithms running on architectures where the flash disks replace the hard disks. Another likely scenario is that rather than replacing hard disk, flash disk may become an additional secondary storage, used together with hard disk. From the algorithm design point of view, it leads to many interesting questions. A fundamental question here is how can we best exploit the comparative advantages of the two devices while running an application algorithm.

The simple idea of directly using external memory algorithms with input and intermediate data randomly striped on the two disks treats both the disks as equal. Since the sequential throughput and the latency for random I/Os of the two devices is likely to be very different, the I/Os of the slower disk can easily become a bottleneck, even with asynchronous I/Os.

The key idea in designing efficient algorithms in such a setting is to restrict the random accesses to a static data-structure. This static data-structure is then kept on the flash disk, thereby exploiting the fast random reads of these devices and avoiding unnecessary writing. The sequential read and write I/Os are all limited to the hard disk.

We illustrate this basic framework with the help of external memory BFS algorithm of Mehlhorn and Meyer [9].

The BFS algorithm of Mehlhorn and Meyer [9] involves a preprocessing phase to restructure the adjacency lists of the graph representation. It groups the nodes of the input graph into disjoint clusters of small diameter and stores the adjacency lists of the nodes in a cluster contiguously on the disk. The key idea is that by spending only one random access (and possibly some sequential accesses depending on cluster size) in order to load the whole

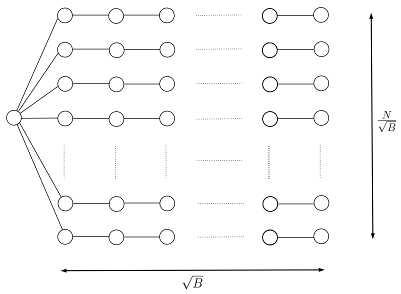


Figure 4.1: A graph class that forces the Mehlhorn/Meyer BFS algorithm to incur its worst case I/O complexity

cluster and then keeping the cluster data in some efficiently accessible data structure (hot pool) until it is all used up, the total amount of I/Os can be reduced by a factor of up to $\sqrt{B}$ on sparse graphs. The neighboring nodes of a BFS level can be computed simply by scanning the hot pool and not the whole graph. Removing the nodes visited in previous two levels by parallel scanning gives the nodes in the next BFS level (a property true only for undirected graphs). Though some edges may be scanned more often in the pool, random I/Os to fetch adjacency lists is considerably reduced.

This algorithm is well suited for our framework as random I/Os are mostly restricted to the data structure keeping the graph clustering, while the hot pool accesses are mostly sequential. Also, the graph clustering is only stored once while the hot pool is modified (read and written) in every iteration. Thus, we keep the graph clustering data structure in the flash disk and the hot pool on the hard disk.

We ran a fast implementation [2] of this algorithm on the graph class shown in Figure 4.1 which is considered difficult for the above mentioned algorithm. This graph class is a tree with $\sqrt{B}+1$ BFS levels. Level 0 contains only the source node which has an edge to all nodes in level 1. Levels $1 \ldots \sqrt{B}$ have $\frac{n}{\sqrt{B}}$ nodes each and the i^{th} node in j^{th} level $(1 < j < \sqrt{B})$ has an edge to the i^{th} node in levels $j-1$ and $j+1$.

As compared to striping the graph as well as pool randomly between the hard disk and the flash disk, the strategy of keeping the graph clustering data structure in flash disk and hot pool in hard disk performs around 25% better. Table 4.1 shows the running time for the second phase of the algorithm for a 2^{28}-node graph. Although the number of I/Os in the two cases are nearly the same, the time spent waiting for I/Os is much better for our disk allocation strategy, leading to better overall runtime.

The cluster size in the BFS algorithm was chosen in a way so as to balance

Operation	Random striping		Our strategy	
	1 Flash + 1 Hard disk	2 Hard disks	Same cluster size	Smaller cluster size
I/O wait time	10.5	6.3	7.1	5.8
Total time	11.7	7.5	8.1	6.3

Table 4.1: Timing (in hours) for the second phase of Mehlhorn/Meyer's BFS algorithm on 2^{28}-node graph

the random reads and sequential I/Os on the hard disks, but now in this new setting, we can reduce the cluster size as the random I/Os are being done much faster by the flash memory. Our experiments suggest that this leads to even further improvements in the runtime of the BFS algorithm.

5 Conclusions and Future work

We have characterized the performance of flash storage devices by benchmarking more than 20 different such devices. We conclude that the read/write/erase behavior of flash is radically different than that of other external block devices like hard disks. Though flash devices have faster random access than the hard disk, they can neither provide the read/write throughput of the disks (the ones that can provide are far more expensive than the same capacity hard disk), nor provide faster random writes than hard disks. We found out that access costs on flash devices also depend on the past history (particularly, the number of random writes done before) and misalignment, but not on the aging of devices.

We also showed that existing RAM model and external memory algorithms can not realize the full potential of the flash devices. Many interesting open problems arise in this context such as how best can one sort (or even search) on a block based device where the read and write costs are significantly different.

Furthermore, we observe that in the setting where the flash becomes an additional level of secondary storage and used together with hard disk rather than replacing it, one can exploit the comparative advantages of both by restricting the random read I/Os to a static data structure stored on the flash and using the hard disk for all other I/Os.

Our results indicate that there is a need for more experimental analysis to find out how the existing external memory and cache-oblivious data structures like priority queues and search trees perform, when running on flash devices. Such experimental studies should eventually lead to a model for predicting realistic performance of algorithms and data structures running on flash devices, as well as on combinations of hard disks and flash devices. Coming up with a model that can capture the essence of flash devices and yet is simple enough to design and analyze algorithms and data structures, remains an important challenge.

As a first model, we may consider a natural extension of the standard external-memory model that will distinguish between block accesses for reading and writing. The I/O cost measure for an algorithm incurring x read I/Os and y write I/Os could be $x + c_W \cdot y$, where the parameter $c_W > 1$ is a penalty factor for write accesses.

An alternative approach might be to assume different block transfer sizes, B_R for reading and B_W for writing, where $B_R < B_W$ and $c_R \cdot x + c_W \cdot y$ (with $c_R, c_W > 1$) would be the modified cost measure.

Appendix A Our benchmarking software and methodology

Our benchmarking software (running under linux) performs a series of experiments on a given block devices according to instructions in an input file. Each line in the input file describes one experiment, which usually consists of many reads or writes. Each experiment can consist of sequential or random reads or writes with a given block size. The accesses can be aligned to a multiple of the block size or misaligned by a given offset. Sequential accesses start at a random multiple of the block size. Random accesses generate and use a permutation of the possible starting addresses (so addresses are not repeated unless the entire address space is written). The line in the input file describes the number of accesses or a time limit. An input line can instruct the program to perform a self scaling experiment [4], in which the block size is repeatedly doubled until the throughput increases by less than 2.5%.

The buffers that are written to flash include either the approximate age of the device (in number of writes) or the values `0x00` to `0xff`, cyclically.

The block device is opened with the `O_DIRECT` flag, to disable kernel caching. We did not use raw I/O access, which eliminates main memory buffer copying by the kernel, because it exhibited significant overheads with small buffers. We assume that these overheads were caused by pinning user-space pages to physical addresses. In any case, buffer copying by the kernel probably does not have a large influence at the throughput of flash memories (we never measured more than 30 MB/s).

We used this program to run a standard series of tests on each device. The first tests measure the performance of aligned reads and writes, both random and sequential, at buffer sizes that start at 512 and double to 8 MB or to the self-scaling limit, whichever comes last. For each buffer size, the experiment starts by sequentially writing the entire device using a 1 MB buffer, followed by sequential reads at the given buffer size, then random reads, then sequential writes, and finally random writes. Each pattern (read/write, sequential/random) is performed 3 times, with a time limit of 30 seconds each (90 seconds total for each pattern).

We also measure the performance of sequential writes following bursts of random writes of varying lengths (5, 30, and 60 seconds). As in the basic test, each such burst-sequential experiment follows a phase of sequentially writing the entire device. We measure and record the performance of the sequential writes at a higher resolution in this test, using 30 phases of 4 seconds each, to assess the speed at which the device recovers from the random writes. We tested random bursts of both 2 KB writes and of random writes at the same buffer size as the subsequent sequential writes.

Finally, we also measure the performance of misaligned random writes. These experiments consisted of 3 phases of 30 seconds for each buffer size and for each misalignment offset.

Entire-device sequential writes which separate different experiments are meant to bring the device to roughly the same state at the beginning of each test. We cannot guarantee that this always returns the logical-to-physical mapping to the same state (it probably does not), but it allows the device some chance to return to a relatively simple mapping.

We also used the program to run endurance tests on a few devices. In these experiments, we alternate between 1000 sequential writes of the entire logical address space and detailed performance tests. In the detailed phases we read and write on the device sequentially and randomly, in all relevant buffer sizes 3 times 30 seconds for each combination. The phases consisting of 1000 writes to the entire address space wear out the device at close to the fastest rate possible, and the detailed experiments record its performance as it wears out.

It is possible that there are other factors that influence performance of some LBA flash devices. However, since many modifications to the benchmarking methodology can be implemented simply by editing a text file, the benchmarking program should remain useful even if more behaviors need to be tested in the future. Of course, some modifications may also require changes to the program itself (e.g., the alignment parameter was added relatively late to the program).

Appendix B Exceptions to the general access patterns of flash memory devices

In most devices, reads were faster than writes in all block sizes. This typical behavior is shown in Figure 3.1a. But as Figure 3.1b shows, this is not a universal behavior of LBA flash devices. In the device whose performance is shown in Figure 3.1b, large sequential writes are faster than large sequential reads. This shows that designers of such devices can trade off read performance and write performance. Optimizing for write performance can make sense for some applications, such as digital photography where write performance can determine the rate at which pictures can be taken. To professional photographers, this is more important than the rate at which pictures can be viewed on camera or downloaded to a computer.

Poor random-write performance is not a sign of poor design, but part of a tradeoff. All the devices that achieve sequential-write performance of over 15 MB/s (on large buffers) took more than 100 ms for small random writes. The two devices with sub-10ms random writes achieved write bandwidths of only 6.9 and 4.4 MB/s. The reason for this behavior appears to be as follows. To achieve high write bandwidths, the device must avoid inefficient erasures (ones that require copying many still-valid pages to a new erase block). The easiest way to ensure that sequential writes are fast is to always map contiguous logical pages to contiguous physical pages within an erase block. That is, if erase blocks contain, say 128 KB, then each contiguous logical 128 KB block is mapped to the pages of one erase block. Under aligned sequential writes, this leads to optimal write throughput. But when the host writes small random blocks, the device performs a read-modify-write of an entire erase block for each write request, to maintain the invariant of the address mapping.

On the other hand, the device can optimize the random-write performance by writing data to any available erased page, enforcing no structure at all on the address mapping. The performance of this scheme depends mostly on the state of the mapping relative to the current access pattern, and on the amount of surplus physical pages. If there are plenty of surplus pages, erasures can be guaranteed to be effective even under a worst-case mapping. Suppose that a device with n physical pages exports only $n/2$ logical pages. When it must erase a block to perform the next write, it contains $n/2$ obsolete pages, so on at least one erase block half the pages are obsolete. This guarantees a 50% erasure effectiveness. If there are only few surplus pages, erasures may free only a single page. But if the current state of the mapping is mostly contiguous within each erase block and the access pattern is also mostly contiguous, erasures are effective and do not require much copying.

This tradeoff spans a factor of 10 or more in random-write performance and a factor of about 4 or 5 in sequential-write performance. System designers selecting an LBA flash device should be aware of this tradeoff, decide what tradeoff their system requires, and choose a device based on benchmark results.

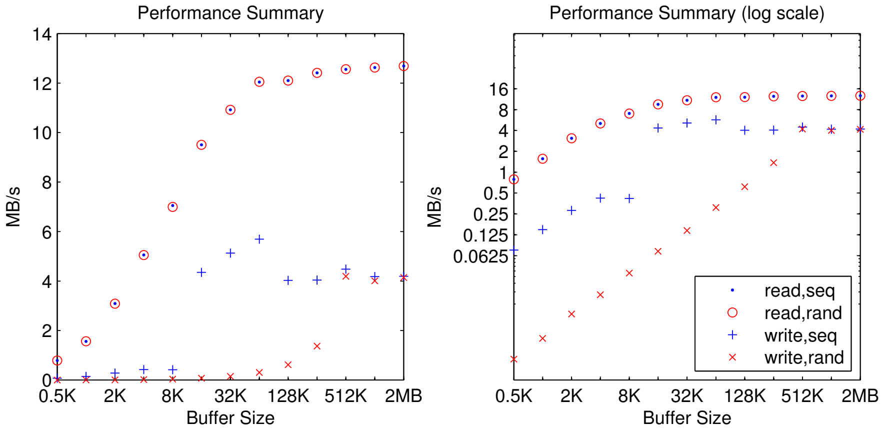


Figure B.1: Speeds of the 512M Toshiba TransMemory USB flash device. This device achieves its maximum write speed at a 64K buffer size.

The read performance does depend on block size, but usually not on whether the access pattern is random or sequential. On a few devices, like the one whose performance is shown in Figure B.1, sequential reads are faster than random reads, but usually the two patterns achieve similar performance.

In most cases, performance increases or stays the same when the block size increases. But Figure B.1 shows an exception. The best sequential-write performance of this occurs with blocks of 64 KB; on larger blocks, performance drops (by more than 20%).

Appendix C Recovery time after a burst of random writes

Figure C.1 presents the performance of a device in which random writes slow down subsequent sequential operations. In these experiments, we performed t seconds of random writing, for $t = 5, 30$ and 60. We then measured the performance of sequential writes during each 4 second period for the next 120 seconds. The two graphs in the middle show the median performance in these 30 4-second periods relative to the steady-state performance of the same pattern (read or write and with the same block size). As we can see, for very small blocks the median performance in the two minutes that follow the random writes can drop by more than a factor of two. Even on larger blocks, performance drops by more than 10%.

The two graphs in the middle row of Figure C.1 differ in the block size during the t seconds of random writes. In the middle-left graph, the random writes were of the same size as the subsequent operation, whereas in the middle-right graph the random writes were always of 2 KB buffers. The behavior of this particular device in the two cases is similar, but on other devices later the two cases differ. When the two cases differ, random writes of 2 KB usually slow down subsequent writes more than random writes of larger blocks. This is typified by the results shown in Figure C.2.

In experiments not reported here we explored the effects of random writes on subsequent read operations and on subsequent random writes. We did not discover any effect on these subsequent operations, so we do not describe the detailed results of these experiments.

The graph on the lower-left corners of Figures C.1 and C.2 show how long it took the device to recover back to 60% of the median performance in the two minutes following the random writes. The device in Figure C.1 usually recovers immediately to this performance level, but in some buffer sizes, it can take it 20-30 seconds to recover. Note that recovery here means a return to a 0.6 fraction of the median post-random performance, not to the base

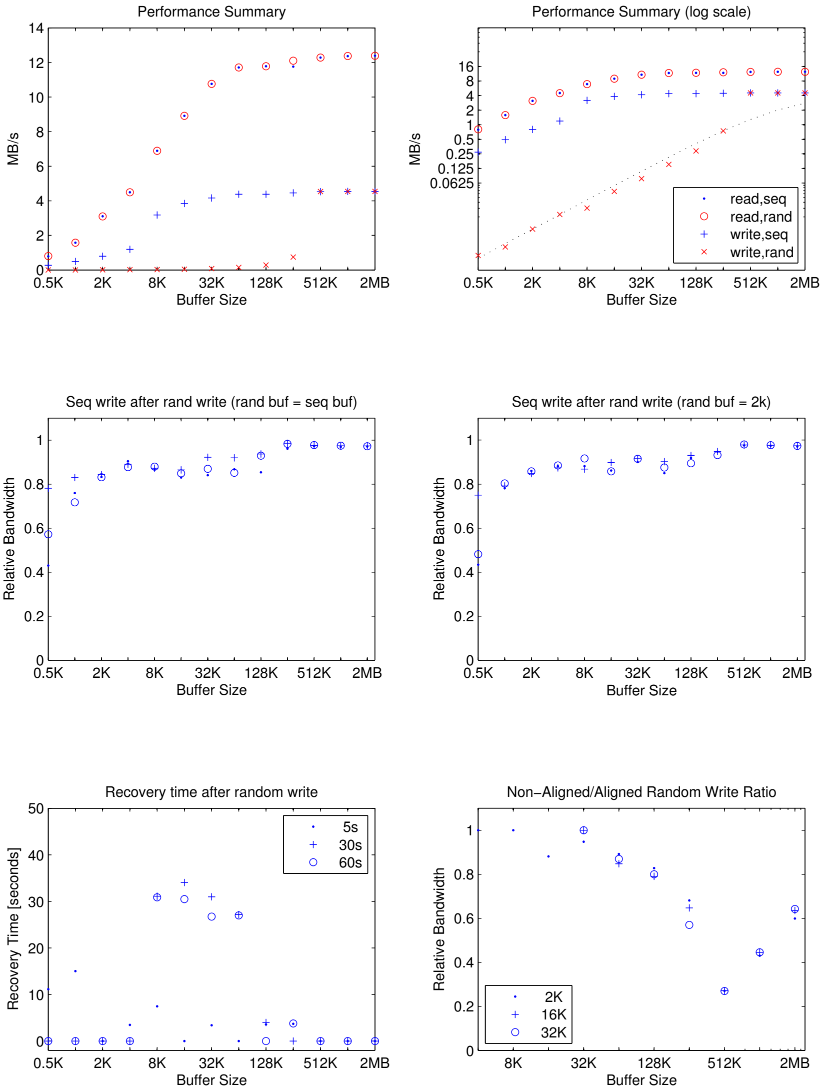


Figure C.1: Toshiba TransMemory USB flash drive results. The top two graphs show the speeds. The two graphs in the middle show how the device is affected by random writes. The bottom left graph shows the time it takes to return back to 60% of the median speed. The bottom right graph shows the effect of misaligned calls on random writes.

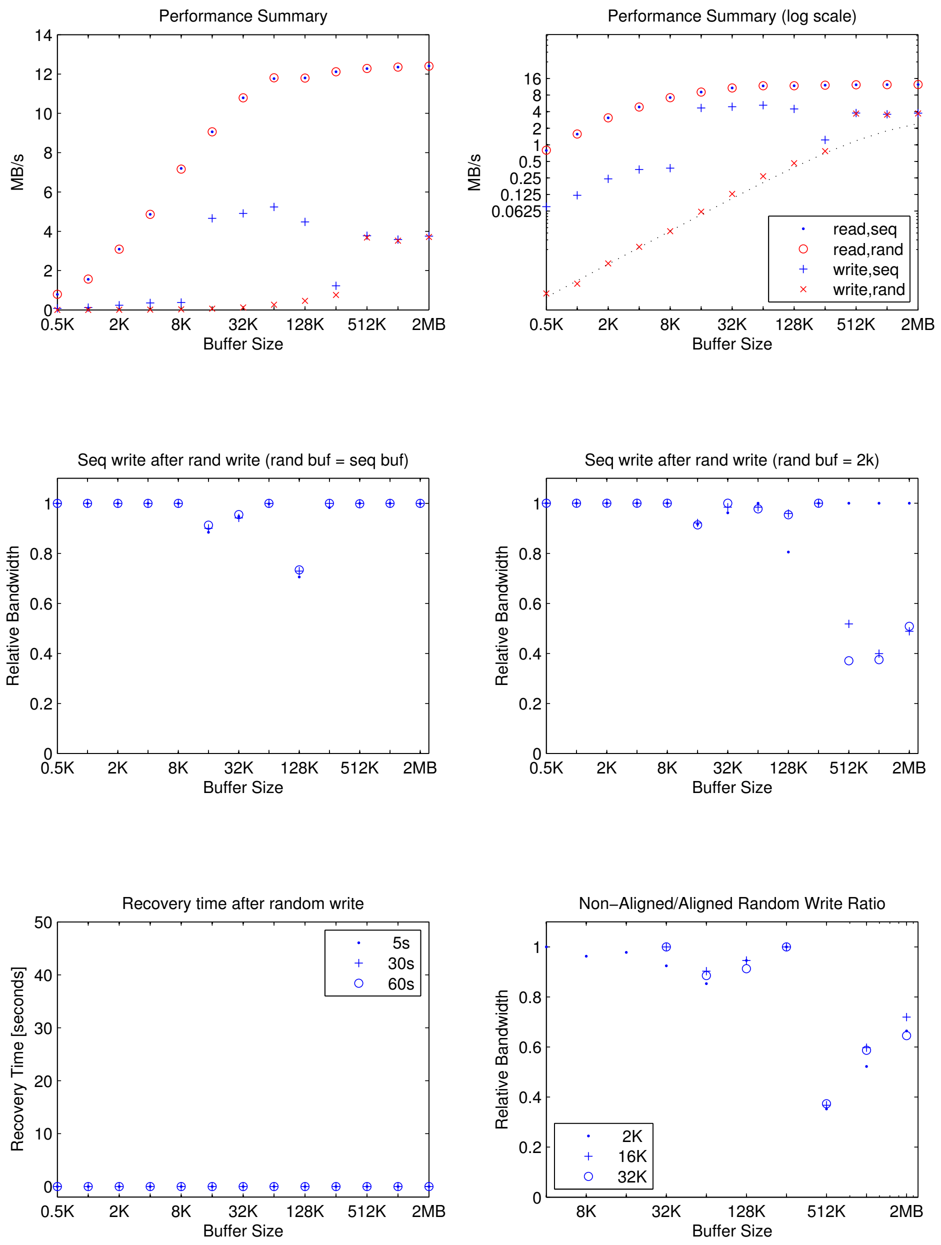


Figure C.2: Results of the M-Systems mDrive 100 USB device, showing a constant decrease in the sequential write speed, with no recovery time.

performance in the particular access pattern.

Figure C.3 presents the recovery time in a different way, on a time line. After a 30 seconds random write time, the speed of the sequential write slows

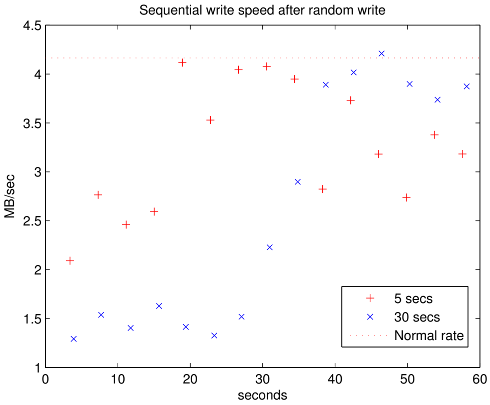


Figure C.3: A time line showing the sequential write performance with 32KB blocks of the device in FigureC.1. The time line starts at the end of 5 or 30 seconds of random writes (again with a 32KB buffer size). The markers show the write bandwidth in each 4-second period following the random writes.

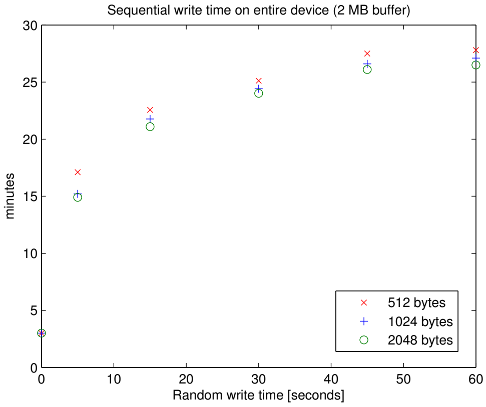


Figure C.4: An example of extreme recovery times, as observed in the 2GB Kingston DT Elite 2.0. The graph shows the time (measured in minutes) it takes to write the entire device sequentially with a 2MB buffer size after random writes of 5 to 60 seconds. Random writes were performed using buffer sizes of at most 2KB.

down to about 30% of the normal speed. After 30 seconds of a sequential write, the speed climbs back towards the normal speed. We have seen similar behaviors in other devices that we tested.

On the high-end 2 GB Kingston DT Elite 2 device, random writes with buffer sizes of 2 KB or less cause a drop in the the performance of subsequent sequential writes to less than 5% of the normal (with the same buffer size). The device did not recover to its normal performance until it was entirely rewritten sequentially. Normally, it takes 3 minutes to write the entire device sequentially with a buffer size of 2 MB, but after random small-buffer writes, it can take more than 25 minutes, a factor of 8 slower (Figure C.4). We observed the same behavior in the 4 GB version of this device.

We have also observed many devices whose performance was not affected at all by random writes.

Bibliography

[1] A. Aggarwal and J. S. Vitter. The input/output complexity of sorting and related problems. In *Communications of the ACM*, volume 31(9), pages 1116–1127, 1988.

[2] D. Ajwani, U. Meyer, and V. Osipov. Improved external memory bfs implementations. In *Proceedings of the ninth workshop on Algorithm Engineering and Experiments ALENEX'07*, pages 3–12, 2007.

[3] A. Birrell, M. Isard, C. Thacker, and T. Wobber. A design for high-performance flash disks. In *SIGOPS Operating Systems Review*, volume 41(2), pages 88–93, 2007.

[4] P. M. Chen and D. A. Patterson. A new approach to I/O performance evaluation—self-scaling I/O benchmarks, predicted I/O performance. In *Proceedings of the 1993 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 1–12, 1993.

[5] M. Frigo, C. E. Leiserson, H. Prokop, and S. Ramachandran. Cache-oblivious algorithms. In *40th Annual Symposium on Foundations of Computer Science FOCS*, pages 285–297. IEEE Computer Society Press, 1999.

[6] E. Gal and S. Toledo. Algorithms and data structures for flash memories. In *ACM Computing Surveys*, volume 37, pages 138–163, 2005.

[7] A. Goldberg and R. Werneck. Computing point-to-point shortest paths from external memory. In *Proceedings of the 7th Workshop on Algorithm Engineering and Experiments (ALENEX'05)*. SIAM, 2005.

[8] S.-W. Lee and B. Moon. Design of flash-based DBMS: an in-page logging approach. In *SIGMOD Conference*, pages 55–66. ACM, 2007.

[9] K. Mehlhorn and U. Meyer. External-memory breadth-first search with sublinear I/O. In *Proc. 10th Ann. European Symposium on Algorithms (ESA)*, volume 2461 of LNCS, pages 723–735. Springer, 2002.

[10] D. Myers and S. Madden. On the use of NAND flash disks in high-performance relational databases. 2007.

[11] C. H. Wu, L. P. Chang, and T. W. Kuo. An efficient B-tree layer for flash-memory storage systems. In *Proceedings of the 9th International Conference on Real-Time and Embedded Computing Systems and Applications (RTCSA)*, February 2003.

[12] C. H. Wu, L. P. Chang, and T. W. Kuo. An efficient R-tree implementation over flash-memory storage systems. In *Proceedings of the eleventh ACM international symposium on Advances in geographic information systems*, pages 17–24. ACM Press, 2003.