FlashVM: Revisiting the Virtual Memory Hierarchy

Mohit Saxena and Michael M. Swift
Department of Computer Sciences
University of Wisconsin-Madison
{*msaxena,swift*}*@cs.wisc.edu*

Abstract

Flash memory is the largest change to storage in recent history. To date, most research has focused on integrating flash as persistent storage in file systems, with little emphasis on virtual memory paging. However, the VM architecture in most of the commodity operating systems is heavily customized for using disks through software layering, request clustering, and prefetching.

We revisit the VM hierarchy in light of flash memory and identify mechanisms that inhibit utilizing its full potential. We find that software overhead for a page fault can be as high as the latency to read a page from flash, and that swap systems are overly tuned towards the characteristics of disks.

Based on this study, we propose a new system design, FlashVM, that pages directly to flash memory, avoids unnecessary disk-based optimizations, and orders page writes to flash memory without any firmware support. With flash prices dropping exponentially and speeds improving, we argue that FlashVM proves to be a much cheaper and faster virtual memory system.

1 Introduction

> ***Tape is Dead, Disk is Tape, Flash is Disk, RAM locality is King.***
>
> – Jim Gray [6]

Flash memory is aggressively following Moore's law: it is cheaper than DRAM and faster than disks. With these trends, research has focused on integrating flash devices as a replacement to disks for storage [1, 22]. We assert, however, that flash also provides the underlying performance and price characteristics to back virtual memory.

While flash-based virtual memory has been previously investigated [10, 12, 17], it has remained under debate. System administrators have relied on anecdotal wisdom and integrated flash disks as swap space to improve the responsiveness of their systems [3]. Others have advocated avoiding flash for swapping due to its limited write endurance [16]. In this paper, we investigate the truth behind these issues and revisit the VM hierarchy in light of flash memory.

Servers can be statically provisioned with enough memory to avoid swapping, but swap performance still matters for laptop and desktop systems. Workloads on these systems vary widely: running too many programs at once or working on a larger-than-normal data set can cause memory pressure and hence swapping.

Based on the price and performance characteristics of flash, we propose FlashVM, a restructured virtual memory system tuned for swapping to flash memory. FlashVM removes software layers impeding performance and gives complete control over swapping to the VM system, rather than splitting it between the VM system and the block subsystem.

We also show how the VM hierarchy in most modern operating systems is overly tuned to the performance characteristics of disks. This has resulted from decades of disk being the *only* option for swapping [4, 14]. Since flash media behavior differs from disks, most of these disk-focused optimizations inhibit the optimal use of these devices. For example, flash devices have large performance differences based on write patterns (sequential is much better than random), but not read patterns.

This paper revisits the Linux virtual memory hierarchy and presents opportunities to improve its performance with flash memory. First, we find that software latencies vary widely and can be as high as the read access latencies of flash memory, and thus must be streamlined. Second, we find that the Linux VM system makes no attempt to optimize write behavior. With experiments on flash and disks, we show that with the current Linux VM system, memory intensive applications can execute from 60% slower to 65% faster with flash memory than with swapping to disk.

Device	Sequential (MB/s)		Random 4K-IO/s	
	Read	Write	Read	Write
HDD	56	45	120-300/s	
USB flash	11.7	4.3	150/s	20/s
SSD	250	170	35K/s	3.3K/s
PCI-e flash	700	600	102K/s	101K/s

Table 1: Hard disk and NAND flash memory characteristics.

2 Flash Memory: Now and Then

NAND flash memory technology has witnessed several big changes in the recent years. Many new manufacturers have joined the race to produce faster and cheaper flash devices [21]. In this section, we give a background on the key flash characteristics that distinguish them from modern hard disks.

Solid-state disks integrate firmware to provide a disk-like interface, such as SATA, on top of flash storage. This firmware, the Flash Translation Layer (FTL) [8], remaps the logical block addresses to physical flash addresses. It also provides wear leveling to increase write endurance. However, this layer is designed for persistent storage, and hence this mapping is also stored in flash. This is unnecessary if using flash for virtual memory, as swap files are inherently temporary and volatile.

Transfer rates: Flash media differs from hard disks in terms of significantly lower random read latencies (0.1ms vs. 8ms), but the write performance has been a weakness for flash devices.

In contrast to disks, there is a wide range of performance for flash devices available in the market today, as shown in Table 1. Inexpensive devices such as USB flash sticks or camera memories offer moderate read bandwidth but have poor random-write performance. Solid-state disks (SSD), with a standard SATA interface, provide much better performance, about 3x better than the fastest hard disks, with 100MB/s sustained transfer rates. This results from intelligent block mapping schemes, parallel I/O accesses to multiple flash chips (similar to RAID) and write buffering [11]. High-end flash drives connected with the PCI-e interconnect interface are even faster, thereby enabling terabytes of virtual memory with speeds nearer to DRAM.

Thus, the raw performance of flash greatly exceeds disk performance today and is likely to improve further. Much like disks, performance still depends on the workload. Design decisions that produce good performance for sequential workloads may not benefit non-sequential ones, and vice-versa [1]. Flash is much faster than disks for reading random blocks. However, random writes are inherently expensive, because flash is slower to write than read, and requires that large blocks be erased first before being overwritten. Thus, workloads for flash should optimize for sequential writes rather than for sequential reads. In this paper, we discuss this specifically for paging traffic in the context of FlashVM.

Write endurance: Unlike disks, flash restricts the total number of writes to each block. Typical flash devices can sustain 100,000 to 1 million overwrites. Thus, 1-10 *petabytes* of data can be written to a 10GB flash over its lifetime. While file systems can greatly reduce write traffic through caching, paging loses its usefulness if pages are not written to storage and hence is less affected by caching.

To analyze whether flash devices have the overwrite capacity for paging workloads, we analyze a three-day block access trace for swapping traffic on a desktop machine running Linux and configured with 700MB of physical memory. We use a pseudo device driver to intercept the block I/O requests to the swap device. Our estimates indicate a write rate of 948 MB per day for a swap partition of 4 GB. With this write rate, and factoring in wear leveling, even a low-end 4 GB flash device with a limit of 100K overwrites can last over 700 years.

Cost: Until recently, flash memory was far more expensive than either disk or DRAM. However, flash memory capacity and cost per unit is following Moore's Law much more aggressively than DRAM. SanDisk will be offering flash devices with prices around $2.5 per GB in capacities of 60,120 and 240GB; by mid of 2009 [7]. In contrast, 1GB of DRAM today costs ten times more. Some vendors like Samsung are even planning to hike DRAM prices [5]. Thus, it may be soon cost effective to populate a system with large quantities of flash rather than DRAM to satisfy memory-intensive workloads.

3 Virtual Memory Management

Virtual memory paging has focused on swapping to disk for decades [4, 14]. As a result, the performance characteristics of disks, such as seek latency, have become ingrained in its design. In particular, we find that three characteristics of disks are assumed: (1) random read access is slow, (2) disk I/O latencies are much higher than other software latencies, and (3) swap devices are integrated as disk storage also being used by file systems. In this section, we describe the virtual memory hierarchy in Linux and show how it is overly tuned to these properties of disks. Paging to disks follows a similar I/O path in other operating systems [19, 15].

Slow random reads: Virtual memory systems prefetch adjacent pages to improve read performance [2, 14]. Even though the Linux VM system makes no effort to allocate virtually contiguous pages together on disk, it

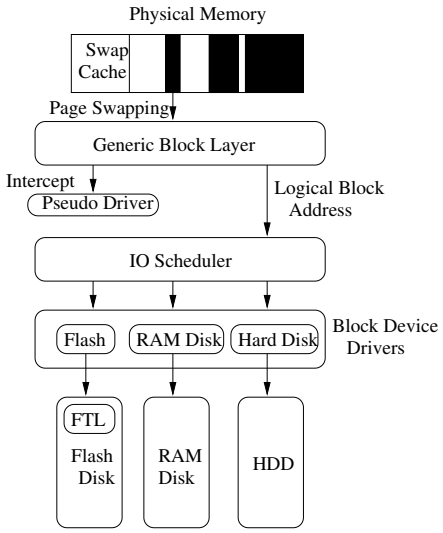

Figure 1: Linux virtual memory hierarchy.

prefetches 8 consecutive pages on disk by default. Thus prefetch, of effectively random pages, is free as the major cost of paging, seeking and rotational latency, must be paid for the first page. However, in the context of flash memory, where random access is cheap, prefetching has little benefit.

Long access latencies: Disks have access times, typically milliseconds, that are much longer than common software latencies. As a result, a paging request can pass through many layers of software without significantly impacting performance.

As shown in Figure 1, a swapped-out page passes through multiple layers in the VM hierarchy. The swap subsystem hands pages to the Generic Block Layer, which is responsible for the conversion of pages into block IO requests, known as *bio* requests in Linux terminology [2].

These *bio* requests are then queued up with the I/O scheduler. The I/O scheduler can reorder, merge or delay a request before passing it to the device driver. A request is delayed to merge it with other contiguous requests that arrive in the near future. This delay can range from 5-6ms [9] to minimize disk seek latencies while reading or writing. Lower in the stack, requests can be delayed or reordered again by the device driver. These additional software layers, which improve disk performance, substantially delay requests. However, flash devices do not suffer the same latencies as disks, so these delays can prove burdensome.

Shared with file system: One reason for these layers is that swap devices are generally shared with file systems. Thus, the OS must provide a common interface for file systems and virtual memory to access. Removing this requirement entails dedicating flash to virtual memory or providing a separate fast-path from the VM system directly to the flash device driver.

4 FlashVM: Revisiting VM Hierarchy

Based on the declining price of flash relative to DRAM and its increasing performance, we propose that future systems be configured with a large amount of dedicated flash to serve as backing store. This can be either attached directly to the system, or a reserved portion of a solid-state disk managed separately. With this configuration, applications with large memory demands can execute much more cheaply than configuring a system with even one-tenth that amount of RAM. Today, flash devices connected over PCI-e can achieve up to 100K random write I/O operations per second at a sustained bandwidth of 600MB/s. As these devices get cheaper, they can be integrated on-board for extending the virtual memory and improve the responsiveness of a system manifold.

However, this design stresses the virtual memory hierarchy, as paging may be far more frequent than in systems today. Hence, we revisit the virtual memory hierarchy and describe the design challenges for FlashVM. We also show why simply using flash devices for extending virtual memory in today's commodity operating systems is not going to tap their full potential.

4.1 VM layering overhead

As shown in Section 3, each swapped out page passes through many layers in the VM hierarchy. This is because the same I/O path below the generic block layer serves both the VM hierarchy and the storage stack. However, the low latency characteristics of flash devices mean that the standard I/O path adds additional overhead for swapping.

We quantify this VM layering overhead by measuring the performance difference between hard and soft page faults to a RAM disk. A soft page fault need not copy data or access the I/O path; it adds a page back to the page table. In contrast, a hard page fault requires copying data from swap storage back into the memory and then adding it to the page table. Thus, the difference between the two faults, less the mandatory cost of copying the page by the RAM disk driver, gives the overhead of the block I/O layer.

We measure these overheads on a 2.5GHz Intel Core 2 Quad with 4GB of RAM. Soft page-fault latency averages 5μs, and copying a 4096 byte buffer takes 10μs (when out of cache). In contrast, the average latency for a hard page fault is 125μs with a significant standard deviation of 6019μs. Thus, the additional overhead associated with VM layering for each hard page fault averages 110μs.

Write Pattern	ext2:HDD		ext2:Flash		nilfs:Flash	
	A	S	A	S	A	S
Seq. (MB/s)	39.4	25.3	20	1.1	16.8	0.63
Random (IO/s)	1,522	120	66	0.1	2,817	149

Table 2: Impact of log-ordering on write performance of flash (A: asynchronous, S: synchronous).

The bulk of this overhead is spent for CPU and I/O scheduling. For a disk, where the minimum read latency exceeds 500μs, these overheads are minor. However, this overhead is *nearly the same as the raw read latency of flash memory*. Thus, software overheads can double the time to swap in a page from a flash device. For FlashVM, the VM system must access flash directly, rather than pass through generic block interfaces, to avoid this layering overhead.

4.2 Ordering Page Writes

As mentioned in Section 2, random writes perform poorly on flash memory. Log-structured file systems [18], or log-structuring in the FTL, improve the performance of flash disks by writing all blocks sequentially. However, the Linux VM system provides no support for swapping to sequential blocks on disk. Rather, the decision of *where* to swap is made independently of *when* to swap, leading to many random writes. As we show below, flash may perform even worse than hard disks for random writes.

We measure the sequential bandwidth and random IO/s for synchronous and asynchronous writes (VM page writes are asynchronous in Linux). We use a 15.8 GB capacity IBM 2.5 inch SSD (Model 43W7617, SATA 1.0 interface), with a random read access latency of 0.2 ms and sustained read bandwidth of 69 MB/s. Table 2 compares the write performance of flash with ext2 and a simple log-structured file system, nilfs [13]) against a 7200 RPM disk with ext2.

At a high level, the hard disk is better at sequential access and the flash device is better at random access. However, the performance depends strongly on the file system layout and whether writes are synchronous or asynchronous. With a conventional file-system layout (ext2), the flash device performs *23 times worse* than the hard disk at asynchronous random writes. However, with nilfs' log structure the flash device performs *85% better* than the disk. Thus, sequential writes are critical to high performance.

In addition, synchronous writes are generally much more expensive for flash. These writes may incur high erase latencies and simultaneous cleaning overheads within the SSD. In contrast, asynchronous writes amortize these costs over a larger group of pages.

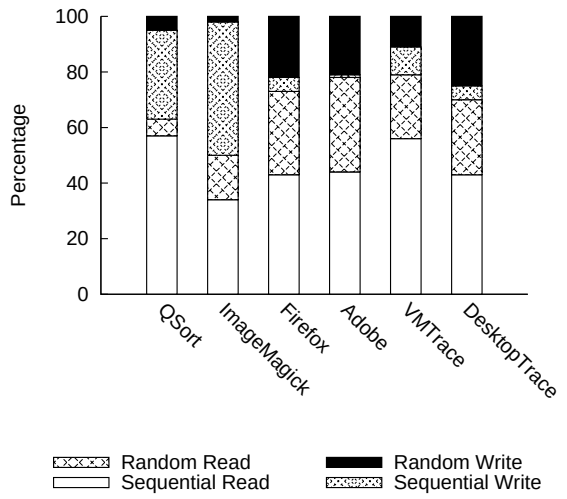

Figure 2: Block access pattern breakdown for different applications.

Thus, the FlashVM system should cluster writes into large sequential groups to optimize for write speed. While high-end flash translation firmware provides this clustering for file systems, they unnecessarily store the virtual-to-physical block address mapping persistently.

To explore these effects on real programs, we investigate the VM access patterns on unmodified Linux kernel 2.6.27 with a 4 GB swap partition on the SSD and 512 MB of physical memory. We profile the swap block-access patterns of 4 applications: a recursive Quick Sort of a large array of random integers, ImageMagick for resizing a large JPEG image, Firefox while surfing the web and streaming videos, and Adobe Reader while viewing pdf files of different sizes. We also trace the swap block accesses on a desktop machine for three consecutive days and on a VMWare virtual machine for 3 hours.

Figure 2 shows the access patterns of these workloads. We characterize as *sequential* those requests submitted to a block device that were adjacent to the previous request; otherwise we consider them *random*. All the access patterns are read-dominated, partially because Linux by default prefetches additional 8 pages on a pagefault.

These traces illustrate two important points about the nature of swapping traffic. First, random reads are common, accounting for more than 20% of the I/O requests from the large programs and the whole-system traces. Flash disks greatly improve performance for these, and hence have the potential to dramatically improve application performance. For example, ImageMagick, with more than 10% random reads, improved execution time on flash by 65% compared to a disk.

More importantly, random writes, which perform poorly on flash, are also common and at times exceed 20% of the accesses in these workloads. As shown in

Table 2, these perform 23 times worse than disk. These random writes contribute QuickSort's poor performance on flash, which is 60% worse than on a disk.

As these results show, simply applying an existing VM system to a flash device, despite better random read performance, may not improve paging performance. Rather, the VM system must be tuned for the particular characteristics of flash, in particular avoidance of random writes.

4.3 Prefetching

As mentioned in Section 3, Linux opportunistically prefetches 8 pages with each swapped-in page. Prefetching for disks is useful to amortize their high seek latency and to overlap I/O with other computation [20]. On the other hand, flash devices provide an order of magnitude lower read access latency and little penalty for random access. Therefore, prefetching functions embedded in the VM system must be updated for flash. We found, for example, that disabling prefetching for ImageMagick improved performance with flash by 27% and for QuickSort by 5%.

5 Conclusions

FlashVM is a promising alternative to existing virtual memory architectures. With flash memory getting cheaper than DRAM and faster than disks, we foresee systems populated with large quantities of flash memory rather than DRAM for satisfying memory-intensive workloads. However, the variable and different performance characteristics of flash memory as compared to disks require the re-design of memory management in modern operating systems. In this paper, we list the primary design challenges for FlashVM. Finally, we show that FlashVM has the potential to provide significant performance improvements for real-world applications.

References

[1] N. Agarwal, V. Prabhakaran, T. Wobber, J. Davis, M. Manasse, and R. Panigrahy. Design tradeoffs for ssd performance. In *USENIX*, 2008.

[2] D. P. Bovet and M. Cesati. *Understanding the Linux Kernel, Third Edition*. O'Reilly Media, Inc., 2005.

[3] Dan's Data. Using flash-swap in Windows Vista. `http://dansdata.com/flashswap.htm`.

[4] E. W. Dijkstra. The structure of the-multiprogramming system. *Communications of the ACM*, 11(5), May 1968.

[5] EE Times, Jan. 2009. `http://www.eetimes.com/news/semi/showArticle.jhtml?articleID=212700554`.

[6] J. Gray. Tape is dead, disk is tape, flash is disk, ram locality is king, Dec. 2006. `http://research.microsoft.com/en-us/um/people/gray/talks/flash_is_good.%ppt`.

[7] InformationWeek. Sandisk Flash Devices. `http://www.informationweek.com/news/storage/portable/showArticle.jhtml?%articleID=212701499&subSection=All+Stories`.

[8] Intel. Understanding the flash translation layer (ftl) specification, Dec. 1998. Application Note AP-684.

[9] S. Iyer and P. Druschel. Anticipatory scheduling: A disk scheduling framework to overcome deceptive idleness in synchronous IO. In *SOSP*, 2001.

[10] D. Jung, J.-S. Kim, S.-Y. Park, J.-U. Park, and J. Lee. Fass: A flash-aware swap system. In *IWSSPS*, 2005.

[11] H. Kim and S. Ahn. Bplru: A buffer management scheme for improving random writes in flash storage. In *USENIX FAST*, 2008.

[12] S. Ko, S. Jun, Y. Ryu, O. Kwon, and K. Koh. A new linux swap system for flash memory storage devices. In *ICCSA*, 2008.

[13] R. Konishi, Y. Amagai, K. Sato, H. Hifumi, S. Kihara, and S. Moriai. The linux implementation of a log-structured file system. *ACM SIGOPS Operating Systems Review*, 40(3), July 2006.

[14] H. M. Levy and P. H. Lipman. Virtual memory management in the VAX/VMS operating system. *Computer*, 15(3):35–41, 1982.

[15] R. McDougall and J. Mauro. *Solaris Internals: Solaris 10 and OpenSolaris Kernel Architecture, Second Edition*. Prentice Hall PTR, 2006.

[16] O. Narasimhan. Optimizing systems to use flash memory as a hard drive replacement. In *Sun BluePrints Online*, 2008.

[17] S. Park, D. Jung, J. Kang, J. Kim, and J. Lee. CFLRU: A replacement algorithm for flash memory. In *CASES*, 2006.

[18] M. Rosenblum and J. K. Ousterhout. The design and implementation of a log-structured file system. *ACM Transactions on Computer Systems*, 10(1), 1992.

[19] M. E. Russinovich and D. A. Solomon. *Microsoft Windows Internals, Fourth Edition*. Microsoft Press, 2005.

[20] E. Shriver, C. Small, and K. A. Smith. Why does file system prefetching work? In *USENIX*, 1999.

[21] SSD-Reviews.com. Solid State Device review. `http://ssd-reviews.com`.

[22] M. Wu and W. Zwaenepoel. envy: A non-volatile, main memory storage system. In *ASPLOS-VI*, 1994.