White Paper

# Using NetApp Flash Cache (PAM II) in Online Transaction Processing

Stephen Daniel, Saad Jafri, NetApp
August 2009 | WP-7082-0809

## FLASH CACHE IMPROVES PERFORMANCE AND STORAGE EFFICIENCY OF ONLINE APPLICATIONS

NetApp Flash Cache (PAM II) is the second generation of intelligent caching modules for FAS family and V-Series storage systems. This paper shows a system that is running at its performance limit and discusses the relative merits of upgrading the system by adding additional disk drives or by adding Flash Cache modules.

TABLE OF CONTENTS

## 1 INTRODUCTION

The storage industry is undergoing a major transformation as flash memory technology transitions from a tool for storing cell phone address books into an enterprise-ready performance acceleration technology for large scale disk storage. This new technology opens a number of options for storage managers who are confronting a need for increased performance.

Consider a storage system that supports a typical online transaction processing (OLTP) workload. Such workloads are a mix of reads and writes, random and sequential accesses. Because rotating disk drives perform poorly under random access workloads, the throughput of OLTP systems often becomes limited by the performance of even the fastest disk drives. Before the advent of inexpensive flash-based storage, the usual path to improve the performance of such systems was to add more disk drives. For some systems and some circumstances, this might still be the correct choice. However, our Flash Cache (PAM II) card offers an innovative, simple, and efficient alternative.

Flash Cache, our second generation intelligent caching module, is in essence a large read cache built using flash-based memory technology. The module comes in two versions, one with 256GB of memory and one with 512GB of memory. Midrange and high-end NetApp® storage controllers can use one or more of these cards to cache data for very rapid access (refer to Flash Cache technical specifications for the most up-to-date support matrix). Our Data ONTAP® storage operating system provides fully automatic, tunable control over the data that will be placed into the Flash Cache module.

With the introduction of this new product, you have a choice of upgrade paths. This paper explores these alternatives, shown in Figure 1.

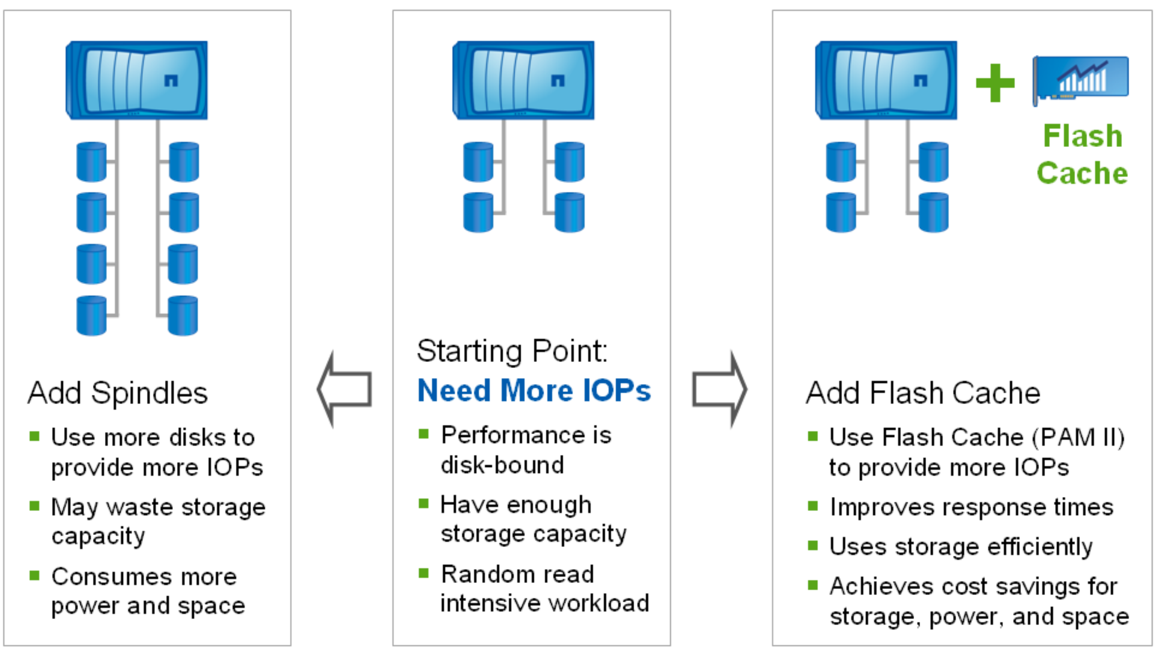


**Figure 1) Upgrade paths**

## 2 OLTP WORKLOADS

OLTP workloads are sometimes over-simplistically classified as random access workloads. To the contrary, OLTP workloads are complex. In-depth examination of running systems shows a more complex mix of work. Papers that have looked at OLTP workloads have typically used very simple emulations such as the IOzone tool to analyze random access performance. Our studies have shown not only that OLTP workloads are complex mixes of random and sequential operations, but also that they have significant locality. This makes sense. Today's data is more frequently accessed than last week's. This month's data is more frequently accessed than last year's.

For purposes of this paper, we have chosen a complex OLTP workload emulation. Our workload is a 40/60 read/write mix and a 60/40 random/sequential mix. Our base block size is 4KB, but the sequential workloads use larger block sizes, up to 64KB. The random access activity covers all of the data set, but accesses are concentrated; 40% of random accesses land within 5% of the data set, and 90% of the random activity accesses only 15% of the data set.

We believe this to be a conservative model of an OLTP workload, because we expect almost all production OLTP applications will cache at least as well as our model. While larger and larger server memories have pushed the read-to-write ratio down, real-world read/write ratios vary widely. We've chosen a write-intensive workload so that we can investigate the utility of a flash-based accelerator in such an environment. Our workload is only moderately concentrated. Anecdotally, we believe that a concentration of 90% of accesses within 5% of the data set would be realistic, but as customer workloads vary we've chosen a more dispersed model.

For our test system we've chosen a data set that is about 15.1TB in size. This data set is larger than many OLTP databases. However, the size was chosen to represent 60% of the total raw storage on the system. We feel that 60% of total raw storage (inclusive of spares and system data) represents a realistic amount of active data on a midrange NetApp storage system.

# 3 TEST SYSTEM

Our baseline test system is a NetApp FAS3160 storage system with two storage controllers and six shelves of 15,000 RPM 300GB high-performance disk drives (total of 84 drives). In its initial configuration, this system can deliver about 31,500 I/O operations per second, using the workload mix described above. The average response time across all of these requests is about 27 milliseconds.

To increase the the configuration's I/O operations per second performance, we then upgrade the baseline system in two distinct ways. One upgrade adds a pair of 256GB Flash Cache (PAM II) cards to each storage controller. The number of disks in this upgraded configuration, 84, matches the baseline configuration. The second upgrade adds an additional 10 shelves of 15,000 RPM 300GB drives (140 disks) to the system instead of the Flash Cache modules.

# 4 RESULTS

Our results show that both upgrade paths are viable. Both can deliver over 55,000 I/O operations per second. However, the resulting capacity, system size, system power, and system response time are all starkly different.

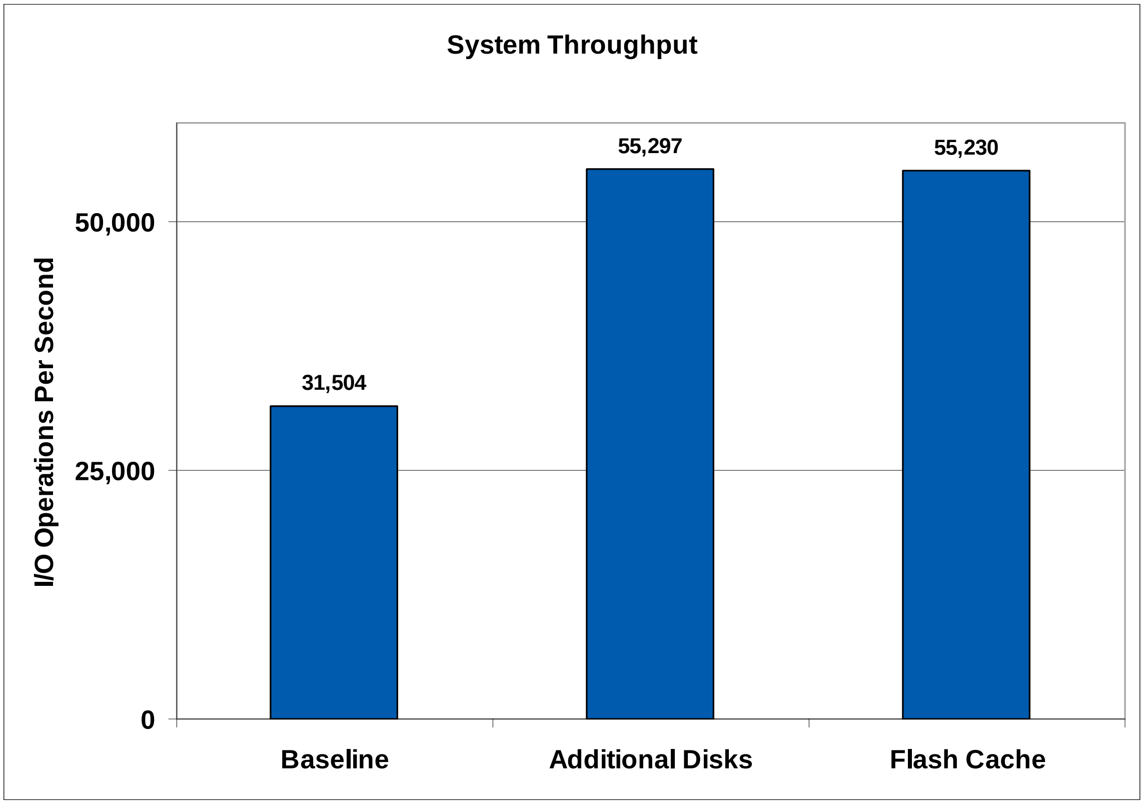


Figure 2) System throughput

Both upgrades produce the same throughput. In essence, both of the upgraded systems are limited by the bandwidth that is available in the FAS3160 storage controllers. However, there are differences. Response time is a key metric that is important to many OLTP systems. Flash memories are much faster than rotating disks at processing random reads. While random reads are a minority of this workload, they do contribute significantly to the response times.

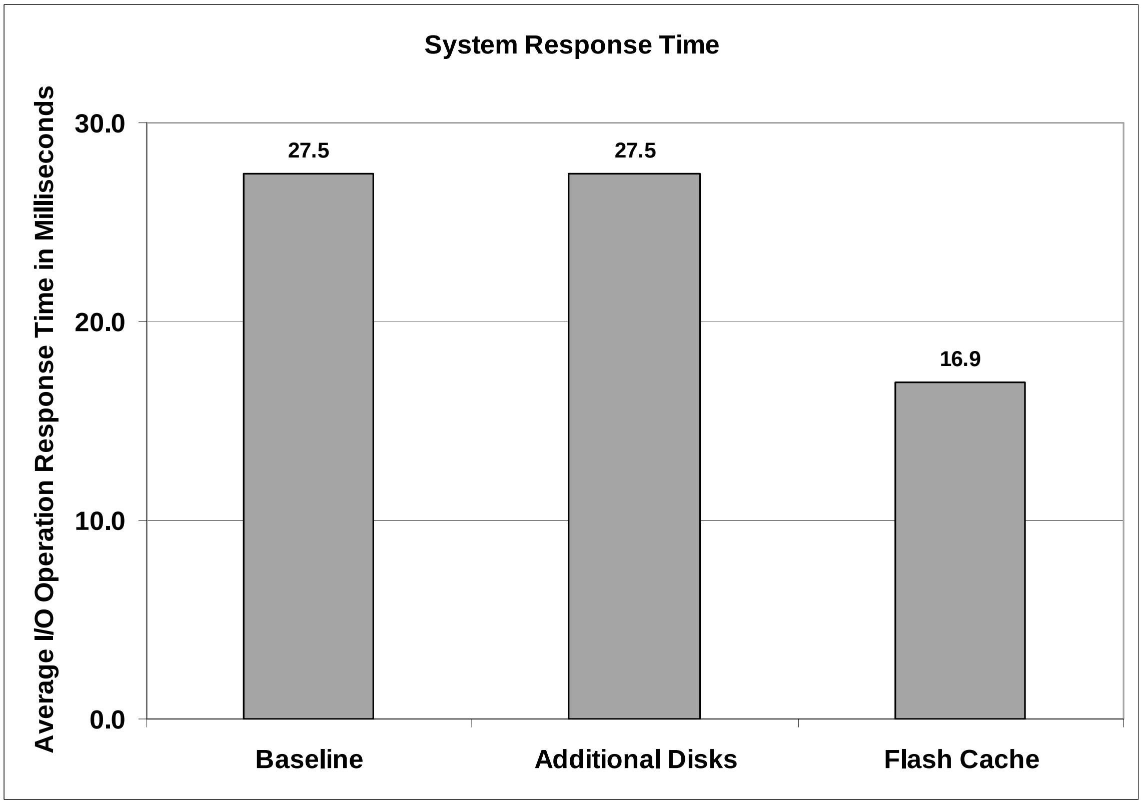


Figure 3) System response time.

As expected, the system configurations that use only rotating disks have approximately the same response time, while using Flash Cache to satisfy the random reads decreases response time by more than one third.

Flash Cache provides this additional performance at a significantly lower price than adding in additional disk shelves. The purchase price of the storage system with Flash Cache is about 30% lower[1] than the purchase price of the system with the upgraded set of disk shelves.

[1] Purchase price calculations made using NetApp list prices for North America in the month of August 2009. Price differences are subject to change over time and might vary by country.

There are also very significant differences in operating costs of the two systems.

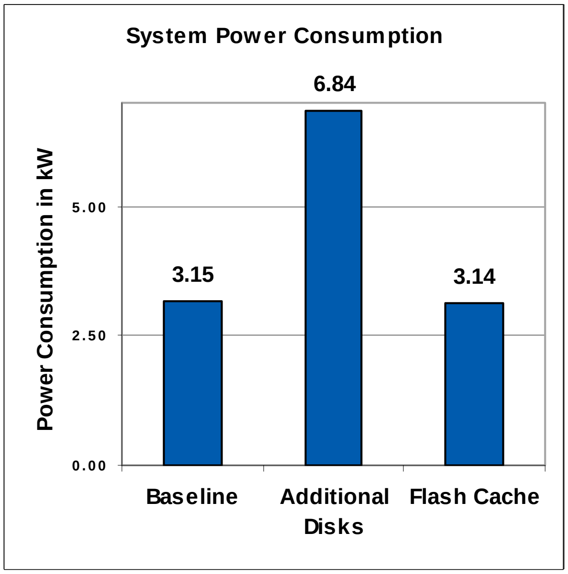


Figure 4) System power consumption

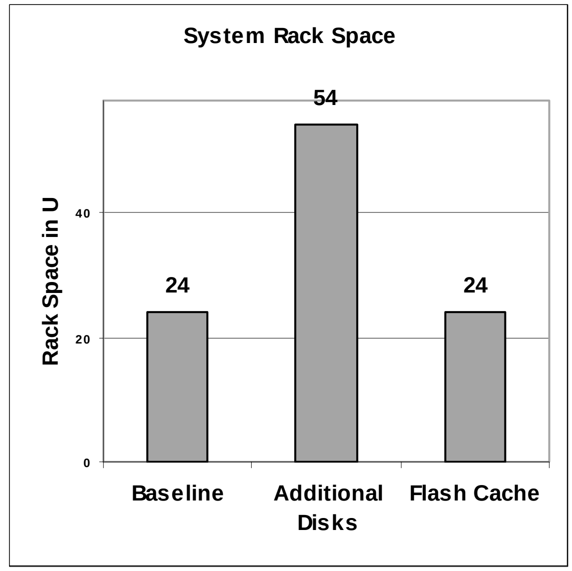


Figure 5) System rack space

As shown, Flash Cache offers significant operational savings. In the systems that are modeled here, improving the performance by adding disks more than doubles the required power, cooling, and rack space. Adding the Flash Cache modules has a negligible impact on power and cooling and requires no additional rack space.

However, there are some advantages to adding disks instead. Clearly the additional disks will add capacity. Whether this capacity is usable depends on the performance profile of the application. If the disks are added only to support the current workload, the use of the new space will have to be limited, or additional usage will impact the current workload.

There is one other advantage to adding the disks. Flash Cache acts as an intelligent read cache. All of the data in cache is secured on disk. Rare events such as a loss of power or failure of the controller hardware will clear the cache. Until the cache warms back up, performance will be degraded. With our workload, this warm up takes about three hours.

# 5 ANALYSIS

The alert reader might notice that the Flash Cache (PAM II) modules nearly doubled the performance of this system, yet the concentrated random reads that are amenable to caching by the Flash Cache cards represent only 30% of the workload. This result is counterintuitive.

The key to understanding the performance results centers on random reads. Random reads are the only workload element that is not already heavily optimized by NetApp storage systems. Data ONTAP already contains a sophisticated read-ahead algorithm to optimize sequential reads. Data ONTAP has long been known for its remarkable ability to optimize random writes, yielding outstanding random write performance even while using relatively large double-parity RAID groups.

The 30% of the workload that the Flash Cache modules accelerate represents the only portion of the workload that is not already accelerated by Data ONTAP. For this reason, Flash Cache has a large impact on the system performance though it only acts on 30% of the incoming requests.

Nearly the same performance improvement could have been achieved by adding enough solid state disks (SSDs). Unfortunately, while SSDs can add significant performance to a disk-limited system, they require careful control over which data will be placed in the SSDs and which will be kept on rotating media. As with Flash Cache modules, SSDs can sharply lower the cost of achieving a performance target, but only if the proper portion of the data set is kept on the SSDs. The Flash Cache modules are managed as a cache, automatically keeping recent data in the flash memory.

## 6 CONCLUSIONS

This paper has shown how to use Flash Cache (PAM II) to improve the performance of a disk-limited system that runs an OLTP workload. The Flash Cache approach offers numerous advantages, including efficiency and response time, over the more common approach of adding high-performance disk drives.

The data in this paper was generated using one of the more powerful NetApp systems, the FAS3160. Flash Cache modules are available across the FAS3100 and FAS6000 families. While the configuration details will vary depending on the controller chosen, we are confident that the savings demonstrated in this paper are available on a variety of workloads and across a wide range of configurations.



www.netapp.com