Dynamic Data Replication on Flash SSD Assisted Video-on-Demand Servers

Ramya Manjunath
Teradata Corporation
ramya.manjunath@teradata.com

Tao Xie
San Diego State University
xie@cs.sdsu.edu

Abstract

Flash memory based SSD (solid state disk) has quick access time, which make it an attractive storage device for read-intensive applications like VoD (Video-on-Demand). However, flash SSD is much more expensive than HDD (hard disk drive). Thus, we propose a hybrid storage architecture that integrates small capacity flash SSDs into HDDs to form a high-performance and cost-effective storage subsystem for VoD servers. Further, on top of the hybrid storage architecture, a dynamic data replication strategy called FLARE (flash for replication) has been developed. Inspired by the insights of a recent user behavior investigation on a large-scale VoD system, FLARE only makes replicas for the first several minutes of a popular video file. A comprehensive simulation study using a real-world trace on a validated disk simulator demonstrates that FLARE consistently performs better than conventional pure HDD based data striping.

1. Introduction

Large-scale VoD systems like PowerInfo [8] normally employ a distributed architecture with more than one million users over multiple cities. VoD systems are data-intensive in nature for two reasons. First of all, the size of a modern video file is generally in the range from hundreds of MBytes to several GBytes. The large size of video files makes it difficult to manage them efficiently in a VoD system. Secondly, a large number of users could send requests to a VoD server from different locations simultaneously. Therefore, a VoD server demands a high-performance storage subsystem that can not only provide a huge capacity in the scale of terabytes but also respond user requests promptly. Currently, rotating based magnetic hard disk drives (HDDs) are dominant building blocks for VoD storage systems. Although they are cost-effective and can provide huge capacity, they are facing several difficulties. For example, while disk capacity has been increasing at a rate of about 60% per year, disk access latency has only been improving about 10% per year [15]. Consequently, new storage devices like NAND flash memory based solid state disk (SSD) that do not have long access delay are greatly desirable.

NAND flash memory based SSD (hereafter, flash SSD) is a semiconductor storage module, which is made of arrays of flash memory elements (also called packages) [5][9]. Each element can have multiple dies that share one serial I/O bus and common control signals [5]. Further, each die contains several identical planes. A plane normally has thousands of blocks and one data register as an I/O buffer [4]. In a Samsung 4GB flash element [4], each block has 64 4-KB pages, which are the basic units of read and write in flash. While reads and writes are page-oriented, erasure can be conducted only at block granularity [5]. A block must be erased before being programmed (written). Flash SSD employs a software component called flash translation layer (FTL) implemented in SSD controller to mimic a HDD so that it can replace a HDD without modifying operating system [5]. Modern FTL generally accomplishes three tasks: mapping logical blocks to physical flash pages, garbage collecting, and wear leveling [1][3].

Flash SSDs provide a new avenue towards high performance, highly reliable, and energy-efficient storage systems as they have the following superior advantages [6]. First, they inherently consume much less energy than mechanical mechanism based HDDs [16]. Second, because of their solid state design they are free of mechanical movements, and thus, have enhanced reliability [17]. Finally, they offer much faster random access by eliminating unnecessary seek time delays and rotation latencies [13]. Unfortunately, compared with HDDs, flash SSDs are much more expensive in terms of dollars per gigabyte. Therefore, integrating small capacity flash SSDs into a VoD system to form a high-performance storage system becomes a feasible solution. Based on our knowledge, this research is the first attempt to combine flash SSDs with HDD based VoD systems.

In this paper, we first propose a hybrid storage architecture for VoD systems. Next, we develop FLARE (flash for replication), a dynamic replication strategy for Video-on-Demand servers. FLARE improves the I/O performance by replicating only the first few minutes of popular videos and storing these on flash SSDs, which have faster read speed. The rationale behind FLARE is two-fold. First, based on the report of a recent study of a

978-1-4673-0009-4/12/$26.00 ©2012 IEEE

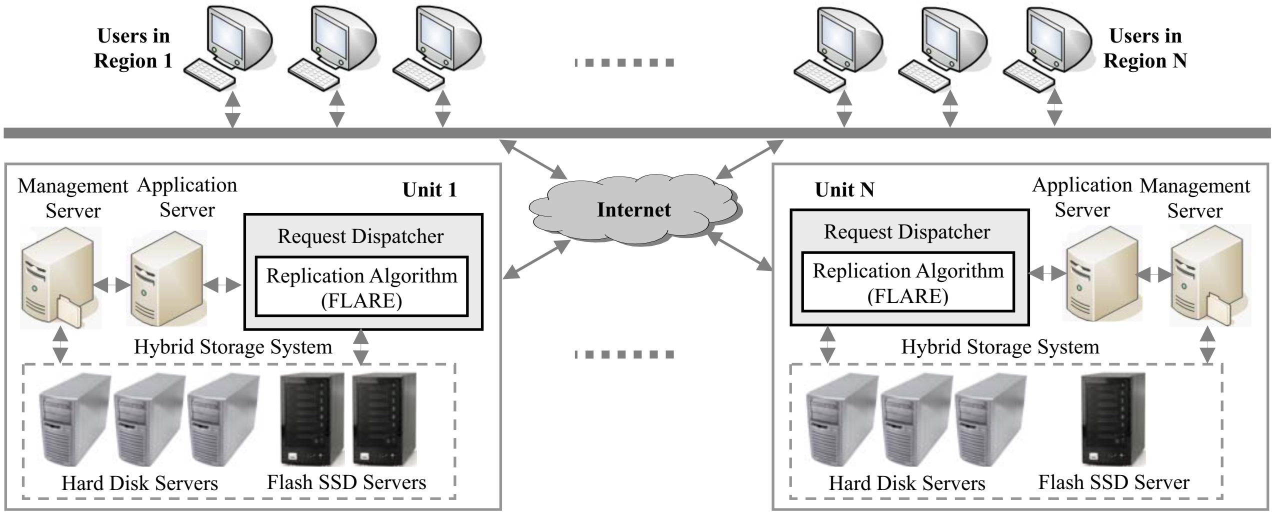

Fig. 1. Architecture of a VoD system.

large VoD system [8], a significant number of users only watch the first several minutes of a selected movie [8]. Therefore, replicas of these video clips can effectively alleviate the burden of original complete video files. Second, replicating only a small part of a large video file can substantially reduce replication cost.

2. Related Work and Motivation

Many efforts have focused on balancing disk load for video-on-demand systems [7][10][11][12]. Three basic mechanisms for load management in video servers include replication, request migration, and striping [7]. When existing copies of a video file are not able to support additional user requests, replication of the video file on another data source becomes necessary. Compared with request migration that requires extra network connection overhead, replication is more effective especially when the cost of hard disk drives has decreased significantly. Disk striping technique stripes videos across multiple HDDs in order to effectively utilize disk bandwidth [12]. Although striping can balance the overall load of the HDDs, it also lowers down the system availability in case of disk failures. The implication is that the degree of striping should be limited to some extent. Therefore, it is still indispensible to create multiple copies of some popular videos to fundamentally improve system performance in terms of response time [10].

A cost-effective approach to building up a scalable VoD system is to integrate a number of VoD servers together in a cluster. Zhou and Xu developed an optimal video replication algorithm and a bounded video placement algorithm in a distributed storage VoD cluster [11]. Their investigations once again demonstrate that replication is an effective way to escalate the performance of a VoD system. A recent study on user behavior of a large VoD system that has more than 150,000 users was reported in [8]. An important observation from [8] is that the majority of sessions (52.55%) are terminated by the user within the first 10 minutes. This evidence suggests that caching the beginning of a large portion of videos can significantly improve user response time [8]. The new discovery provided in [8] motivate us to design a video replication strategy on top of a SSD-HDD hybrid storage architecture such that only the first 10 minutes of popular videos are replicated on flash SSDs to serve at least 50% of all user sessions.

3. The FLARE Strategy

3.1 VoD Architecture

The architecture of the proposed hybrid storage system for a large-scale VoD system is illustrated in Fig. 1. Customers are divided into regional networks and each regional network has one or more server clusters known as units [8]. Servers in each VOD unit cluster can be categorized into three different groups: application servers, management servers, and storage servers. In each unit, one application server caches video metadata, performs authentication, and interacts directly with users to schedule streaming requests. The FLARE strategy is running on the application server. All user requests will be eventually forwarded to one or multiple storage servers, which stream video directly to the user. The application server also balances tasks appropriately across storage servers. A management server monitors all servers in the VOD unit and performs accounting functions. User access pattern is gathered to determine the optimal replication for each individual video file [8]. The storage system of unit 1 has two flash SSD severs and three hard disk servers (see Fig. 1). Each flash SSD server has a flash SSD array, whereas each hard disk sever maintains a hard disk array.

3.2 Algorithm Description

In the proposed VoD server architecture, the videos are primarily stored on the hard disks and the replicas of the popular videos are store on flash SSDs. In FLARE, read requests for the first ten minutes of popular videos that have been replicated will be served by flash SSD. Fig. 2 shows the algorithm of FLARE.

```
Sort requests in ascending order of files.
For every requested file, $f_i$
    $c_i$ = Number of requests to $f_i$ during the last $N$ requests.
    $h_i$ = Number of times the block was accessed / n.
    $Heat\ (f_i, :) = h_i$ //heat of every accessed file.
end for
For every $f_j$ in $Heat$
    if $h_j > HEAT\_MAX$
        $b_j$ = PAGE_SIZE * $f_j$
        $f_j = (h_i, rep\_unit, b_j, :)$
        $rep\_file(:) = f_j$ //Replicate file $f_j$..
    end if
    if $h_j < HEAT\_MIN$ && $replicated\_file(:) == f_j$
        $replicated\_file(j,:)$=[]// Garbage cleaning
    end if
end for

Replicate()
Sort $f_i$ in $rep\_file(:)$ in the decreasing order of $h_i$
For $f_i$ in sorted $rep\_file$
    Get the $no\_of\_replica$.
    Sort disks in the ascending order of $sh_j$.
    Get the disk, $sd_j$ with minimum heat.
    $cnt = 0$.
    while ($cnt < no\_of\_replica$ && $s_i > rep\_unit$ )
        if ( ! $replicated\_file\ (f_i, :, sd_i)$ && $s_i$ > rep_unit)
            $t_i$ = current time
            $req = (b_i, rep\_size, t_i, sd_i)$// SSD write request
            $replicated\_file\ (f_i, :, sd_i) = b_i$
            $sh_j = sh_j + h_i$ //Add heat of the file to disk heat
            $s_i = s_i - rep\_size$ //Add the size of a 10 min video
            $cnt = cnt + 1$
            Sort disks based on the ascending order of $sh_j$.
            Select $sd_j$ with minimum heat.
        else
            Select the next disk from the sorted list.
        end if
    end while
end for
```

Fig. 2. The FLARE strategy.

The number of HDD severs in the hybrid storage system is much larger than that of flash SSD servers. The flash SSD severs and HDD servers together form a cluster. Each server contains disks arranged in an array. Initially, it is assumed that all videos are placed on the hard disks. FLARE replicates the first few minutes of a popular video and places it on the coolest flash SSD. Each Flash SSD is of a fixed size. It is also assumed that there are enough disks on the SSD node to hold the replicas created. The heat of a disk is the sum total of the heat of the files placed on that disk. The entire video will always be available on the hard disk drive. The request dispatcher maintains a record for the replicas generated. If a replica is available for the requested file, then the first few minutes of the request is served faster by the flash SSD. As mentioned earlier, nearly half the requests are terminated within first 10 minutes. In FLARE, read requests for the first ten minutes of popular videos that have been replicated will be served by flash SSD. Since the read requests served by the solid state drives are faster when compared to a hard disk drive, we propose an algorithm where short requests are served faster by the SSD node instead of the traditional HDD node. However, it is important to replicate only a portion of the video file since a video can be of several hours in length. Random writes in flash SSDs can be very expensive. Also, the heavy video files will take many more disks to store the replicas. Thus, to reduce the cost of replication, both in terms of random write costs and the cost of flash SSDs, only a few minutes of popular videos are replicated.

4. Performance Evaluation

In this section, we evaluate the performance of FLARE by comparing it with the conventional disk striping strategy for Video-on-Demand servers. A real-world VoD server trace [8] and a validated simulator toolkit DiskSim [14] plus the SSD model [2] have been utilized in the experiments. The SSD model [2] is a package that extends the DiskSim [14] to provide basic support for solid-state- The flash SSD used in this simulation is a typical 4 GByte device, the Samsung K9W8G08U1M [4]. A hard disk drive node consists of HDD arrays. In our simulation, the disk arrays are modeled using HP C2249A. The parameters for this disk are in a model file provided by DiskSim [14]. FLARE needs at least one SSD node and one HDD node.

The flash SSDs are arranged in an array in an SSD node. Various SSD parameters can be modified. Page size, number of disks used, number of planes are a few to name. For the simulator to function, it is essential to set the right parameters in the parameter file. The output of the request dispatcher is fed into the DiskSim simulator. The simulation output is logged into an output file in a specified location. The output carries in detail the mean response time of the individual component and the overall I/O system. It also gives the read response time and the write response time for the I/O system. It also gives us the total number of requests handled by the system, the system idle time and the average queue length.

To evaluate the performance of FLARE in different levels of workload intensity, we varied the number of input requests. The results were observed by increasing the request size as shown in Fig. 3. One can see that an

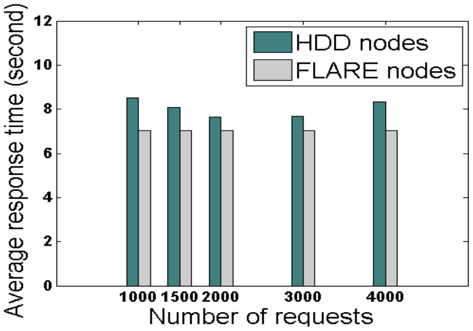

Fig. 3. The impact of number of requests.

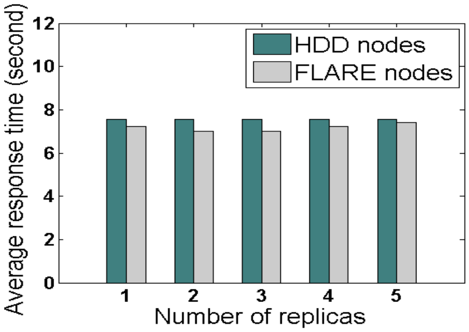

Fig. 4. The impact of number of replicas.

increase in the number of requests greatly improves the performance of the requests. The difference between the total time taken to serve the requests by FLARE and the total time taken to serve the requests by sequentially striped HDD nodes increases with the increase in number of nodes. For a request size of 4000, FLARE shows 23% decrease in the total time taken to serve when compared to the time taken by the HDD algorithm.

The performance of the system deteriorates if many writes are issued to the SSD node. This can be observed in the Fig. 4. Ten minutes of a video file occupies more than 3 MByte of data on the disk. The total time for the I/O system to serve the requests on nodes where FLARE is used, decreases first and then increases. This shows that creating replica adds extra overhead to the system. An optimal number of replicas distribute the workload among the SSD and HDD nodes, thereby balancing the load. However, if the number of replicas of a file is more than what is used by the workload, then creating the extra copies of the same file is an overhead for the system. This issue is aggravated in the case of solid state drives since random writes are expensive in these disks. Hence, it is essential to create an optimal number of replicas, to ensure that the SSD write requests do not deteriorate the performance of the system. The number of replicas to be created depends on the nature of the workload. If a file is being requested frequently, it is useful to create a number of replicas for this popular file.

Varying the load on the SSD node of the system, we observed the behavior of HDD systems as shown in Fig. 5. To analyze the effect of upper threshold heat value we varied *HEAT_MAX* and observed the performance of the HDD node on the system. The simulated system has one SSD node to store short duration replicas of video files and HDD nodes to store the complete videos. Since heat is a measure of popularity of the file an increase in the threshold heat value, increases the number of requests to SSD nodes and thereby increases the total response time of the HDD nodes.

Fig. 6 shows the variation in the performance of FLARE with respect to the conventional striped strategy on a HDD system for varying lengths of the replicas also referred to as replication time. The performance of FLARE improves first and then remains almost a constant when longer lengths of the videos are replicated. To reduce the write costs on an SSD it is important to manage the length of the replicated video. Longer lengths of the replicas can deteriorate the performance as they add redundant writes to the system. It can be observed that the deterioration in performance is not as rapid as that observed when the number of replica is varied. The difference between the total response time for FLARE and conventional algorithm is the maximum when the replication time is twelve minutes. Beyond this the performance decreases first and then remains a constant. This means that beyond twelve minutes the SSD node is losing out on its faster read benefits and the disadvantages of slower writes are creeping in.

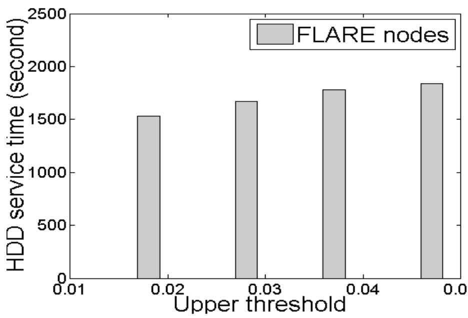

Fig. 5. The impact of upper threshold values.

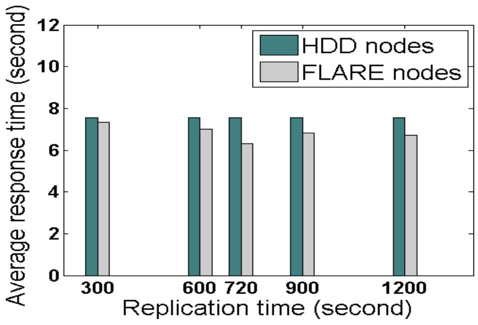

Fig. 6. The impact of replication time.

5. Conclusions

In this paper, we described the design and implementation of FLARE, a replication algorithm on a hybrid storage architecture for a large-scale Video-on-Demand system. We compared it with conventional hard disk drive system where data was striped and placed on the disks in a round robin manner. We showed that by introducing a few small capacity flash SSDs we can build a cost-effective system that achieves high-performance in terms of user response time. This conclusion was verified by an experimental study using DiskSim simulator [14] and a real-world VoD trace from PowerInfo. Simulation results show that using a flash SSD array to store the first 10 minutes of popular video files greatly improves the system performance. We believe that read-intensive workloads like VoD traces can exhibit noticeable performance improvement when FLARE was applied to a hybrid storage architecture. In particular, FLARE improved the performance of the I/O system by up to 23% when compared with the traditional striped HDD system. Moreover, FLARE consistently performs better when compared to the conventional HDD disk striping technique. Based on our knowledge, this research is the first investigation on integrating enterprise flash SSDs into modern large-scale VoD systems.

In future, this work can be extended to show that the energy can also be optimized by using flash SSD nodes along with HDD nodes. The heat or popularity of a video file can decide the degree of replication to further optimize the number of replica that will be created, FLARE can also be optimized to take into account some popularity factors such as a new video added to the system is requested more often that an old one. Including these factors in determining the heat of a file may further improve the performance of the system. This hybrid architecture can also be extended to different types of workloads. Small files with very frequent read requests may improve the performances significantly.

Acknowledgements

This work was supported by the US National Science Foundation under grant number CNS-0834466.

References

[1] "Increasing Flash SSD Reliability." Silicon Systems, Apr 2005. http://www.storagesearch.com/siliconsys-art1.html (accessed May 2009).

[2] "SSD Extension for DiskSim Simulation Environment." Microsoft Research, Microsoft Corporation, 2009. http://research.microsoft.com/en-us/downloads/b41019e2-1d2b-44d8-b512-ba35ab814cd4/.

[3] "Technical Note-29-42: Wear-Leveling Techniques in NAND Flash Devices." Micron Technology, Inc., 2008. http://download.micron.com/pdf/technotes/nand/tn2942_nand_wear_leveling.pdf.

[4] Samsung Electronics. "512M x 8Bit / 1G x 8Bit NAND Flash Memory. K9W8G081M/K9K4G08U0M Flash Memory Datasheet." Samsung Datasheet Catalog. http://www.datasheetcatalog.com/datasheets_pdf/K/9/K/4/K9K4G08U0M.shtml, (accessed July 2009).

[5] Agrawal, N., V. Prabhakaran, T. Wobber, J. D. Davis, M. Manasse, and R. Panigrahy. Proceedings of the USENIX Annual Technical Conference (USENIX '08), June 2008: Design Tradeoffs for SSD Performance. Boston: The Advanced Computing Systems Association.

[6] Weiss, Aaron. "Data Storage the Next Generation." Gadget crazy! Mobile Devices Reshape Our Lives 11, no. 3 (2007).

[7] Venkatasubramanian, N., and S. Ramanathan. 1997. Proceedings of the 17th International Conference on Distributed Computing, Systems (ICDCS '97). May 27-30, 1997: Load Management in Distributed Video Servers.Los Alamitos, CA: IEEE Computer Society Press.

[8] Yu, Hongliang, Dongdong Zheng, Ben Y. Zhao, and Weimin Zheng. 2006. Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems April 18-21, 2006: Understanding User Behavior in Large-Scale Video-on-Demand Systems.

[9] News Release: BiTMICRO stirs up the SSD industry with 416 GB 2.5-inch SATA solid state drive, http://www.bitmicro.com/press_news_releases_20071023.

[10] Wolf, J.L., Yu, P.S., and Shachnai, H. 1997: Disk load balancing for video-on-demand systems. Multimedia Systems, 5:358-370.

[11] Zhou, X., and C. Z. Xu. 2002. Proceedings of the 2002 IEEE 31st International Conference Parallel Processing (ICPP), August 20-23, 2002: Optimal Video Replication and Placement on a Cluster of Video-on-Demand Servers. Vancouver: IEEE Computer Society.

[12] Ozden, B., R. Rastogi, and A. Silberschatz. 1996. International Conference on Multimedia Computing and Systems (ICMCS'96), June 17-23 1996: Disk striping in video server environments, ICMCS. Hiroshima, Japan: IEEE Computer Society.

[13] Bouganim, L. B. Jonsson, and P. Bonnet. 2009. In Proceedings of the 4th Biennial Conference on Innovative Data Systems Research (CIDR'09), January 4-7, 2009: uFLIP: Understanding Flash IO Patterns. Asilomar, California: UC Berkeley.

[14] Bucy, J. S., G. R. Ganger, et al. The DiskSim Simulation Environment Version 4.0 Reference Manual. Pittsburgh: Carnegie Mellon University, 2003.

[15] X. Yu, "Trading capacity for performance in disk arrays," Ph.D. Dissertation, Princeton University, 2004.

[16] K. Cash, "Flash Solid State Disks - Inferior Technology or Closet Superstar?", BiTMICRO Networks, http://www.storagesearch.com/bitmicro-art1.html.

[17] T. Xie and Y. Sun, "PEARL: Performance, Energy, and Reliability Balanced Dynamic Data Redistribution for Next Generation Disk Arrays," Proc. 16th Annual Meeting of the IEEE Int'l Sym. Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), 2008.