Energy-efficient High-performance Storage System

Jun Wang
University of Central Florida
juwang@mail.ucf.edu

Abstract

This project is developing extended versions of RAIDs with low-power, and exploring novel energy-efficient disk array architectures using coding techniques for data intensive computing.

1. Introduction

The energy-related budget has accounted for a large portion of storage system total cost of ownership. It is reported that power and cooling equipment together with electricity represent a significant portion of a data center’s cost — up to 63 percent of the total cost of ownership of its physical IT infrastructure. As of today, there are yet no good solutions to conserve energy for conventional high-speed disk based storage systems, e.g., in server environment. There is a need to invent new ways to address this open question.

There are several major challenges in developing energy-efficient high-performance storage systems.

• Long disk spin-up latency: The power consumed by a commodity disk depends on its current state: active, idle or standby. Spinning a disk down to the standby state can significantly reduce energy consumption. However, in practice conventional high-performance hard drives like SCSI disks cannot be placed in standby to save energy because of the extremely long latencies (of the order of ten seconds) required to transition into and outside the sleep state. The benefit of spinning a disk down to the standby state must be greater than the time and energy costs needed to spin the disk down and then spin it up.

• Dynamic workloads: Server workloads are known to be very bursty and dynamically changing in their access characteristics. This provides both an opportunity and a challenge for energy management techniques to dynamically adapt the energy consumption to the system load. The granularity of the load fluctuations makes the task challenging: rapid adjustment to transient conditions may end up consuming more energy than a nominal amount of over provisioning, while delayed response can result in both lost opportunities for energy savings and unacceptable response times for I/O requests.

• Runtime Control: Scaling a storage system to thousand or more disks while maintaining high-performance and reliability is well-recognized as an extremely challenging problem. Adding the Frequent transitioning to and from a standby state can also affect reliability additional dimension of dynamic energy efficiency increases the complexity considerably. Sophisticated strategies for both short-term and medium-term workload prediction, control algorithms, and analysis techniques are necessary to design a robust and scalable solution.

To deal with the above-mentioned challenges while existing solutions do no address or not well, we have developed dynamic I/O transformation algorithms and novel hierarchical cache software management schemes using existing simple redundancy in disk arrays that serve as a foundation for constructing energy-efficient high-performance storage arrays. We are developing two energy-efficient building-block prototype systems – mirrored array and parity array. We will be developing novel disk array architectures using sophisticated coded redundancy and extend solutions to the distributed environment.

2. Solutions for Parity based Storage System

2.1 Motivation

The long disk spin-up delay derives from a passive spin-up issue (a standby disk has to be spun up to serve incoming requests). In conventional disk
based storage systems, there are three sources of passive spin-ups: i) Non-blocking Read requests that cannot be deferred in a server environment; ii) Derivative Read due to parity updates for write requests in parity disk array, which cannot be deferred for data reliability (Derivative Read is defined as those read requests incurred by parity update in parity disk array); iii) dirty block flushing to enforce coherency between cache and disks.

Zhu et al. thoroughly investigated the impacts of the third type of passive spin-up with alternative write-back and write-through policies [1]. Most of the aforementioned load concentration techniques [2], [3], [4] partially handle the first type of passive spin-up. However they rely on the knowledge of access patterns, especially the requests’ arrival time. Since no prediction algorithm really works well to predict the arrival time, the following scenario could potentially introduce non-negligible overhead: if a predicted arrival time leads to an idle period exceeding the break-even time and the disk is spun down accordingly, but the actual idle period is shorter than the break-even time, the incurred performance penalty overhead is prohibitive, and the energy cost of the disk spin-down and spin-up can be more than that without spin-down. To the best of our knowledge, no one has studied the impact of the second type of passive spin-up on energy consumption. In this paper, we seek novel solutions to deal with the first two types of passive spin-ups, i.e. Non-blocking Read and Derivative Read.

Both Disk level (RAID) and Memory level (Cache) redundancy can be exploited to avoid passive spin-up for conventional disk-based storage system. However, most current solutions rely on DRPM (Dynamic Rotations per Minute) [5] disks in which the passive spin-up is not as severe as that for conventional disk-based system. For example, we observed that additional disk accesses resulting from energy conservation schemes in EERAID (disk level redundancy) can lead to non-negligible performance degradation. To the best of our knowledge, no work has been done to avoid passive spin-up based on memory level redundancy. A recent work called PARAIMD [6] developed a static layout solution to modulate the disk provision according to the load fluctuation for RAID-0/5. It adopted a conventional disk based array model while only worked for light loads changing at a daily scale.

These observations motivate us to reorganize multi-level caches in storage systems such that they can collaborate actively with each other with awareness of the redundancy both in Memory (Storage Cache and Controller Cache) and on Disks (RAID).

The objective is to solve the passive spin-up problem resulting from non-blocking read and derivative read for conventional disk based storage systems, thus achieving both better energy savings and performance.

2.2 Redundancy-based, two-level collaborative I/O cache architecture (RIMAC)

As a small-scale, striped disk array with parity is one of the most important building blocks in modern storage systems, we choose small-scale RAID-5 to study. We accomplish the following outcomes.

• We design a redundancy-based, two-level collaborative I/O cache architecture called RIMAC. The key is to provide an effective collaboration mechanism between storage cache and RAID controller cache, which facilitates redundancy exploitation during power-aware request transformations (see definition below) both in two-level I/O caches and on physical disks.

• Based on the RIMAC architecture, we develop two power-aware read request transformation schemes — Transformable Read in Cache (TRC) and Transformable Read on Disk (TRD), and a power-aware write request transformation policy for parity update to attack the passive spin-up problem. By applying these request transformation schemes, all non-blocking reads and derivative reads to the standby disk can be satisfied either directly by a hit in the storage cache, or indirectly by performing an on-the-fly XOR calculation with i) data from the storage cache, ii) parity from the controller cache, and iii) data or parity from non-standby disk(s) only if needed. The bottom line is to construct the requested data via accessing active disk(s). As a consequence, the low power disks can stay in the standby state for a longer period of time for better energy efficiency. Better performance is achieved by avoiding the long delay of passive spin-up and unnecessary disk accesses.

• We develop a second-chance parity cache replacement algorithm to improve the success rate of the power-aware request transformation policies (e.g. TRC, TRD).

In our trace-driven experiments, we chose three real-life storage server traces: cello99 collected by the HP Storage Research Lab, OLTP trace collected by UIUC Opera group and the search engine trace collected by the Storage Performance Council to thoroughly evaluate RIMAC. Comprehensive simulation results show that RIMAC outperforms existing threshold-based solutions by up to 18% in terms of energy
savings and up to 34% in terms of average response time in RAID-5 systems with less or equal to 8 disks. More detail can be found in our publications [7].

3. Solutions with Dynamic Control for Both Replication and Parity System

3.1 Motivation

Currently, there are two approaches to unbalance workloads for energy saving of conventional disks: relocating data and redirecting requests. Two representative works for relocating data are Massive Array of Idle Disks (MAID) [2] and Popular Data Concentration (PDC) [3]. However, Pinheiro et al. showed that MAID and PDC can conserve energy for conventional disks only when the load on the server is extremely low though they perform much better with multi-speed disks.

The basic idea of redirecting request is to intentionally bypass a data target (for example, a disk) by redirecting requests to other data target(s) to provide the same information to users. We investigated how to use this approach in RAID systems in our previous research works [4] and [7]. Recently, Pinheiro et al. [8] introduced Diverted Accesses to leverage the redundancy in a wide spectrum of storage system for energy conservation. In all of the above mentioned works, the performance impact of redirecting request is not properly dynamically controlled. Literature [9] is among the first to develop several control algorithms to ensure performance guarantees for a multi-speed disk based storage system. But their performance prediction scheme for disks is still in an early stage.

There are several limitations with current server disk energy conservation policies. First, most current energy saving policies for server side storage systems require online system reconfigurations, which means changing the system organization. However, there is no comprehensive study on the performance model or energy model that can be used in the above-mentioned environment. Second, According to our knowledge, current disk power management algorithms only take one performance metric into consideration. However, energy saving policies may degrade at different extents for both throughput and response time, but they are not always proportional. Third, the effects of synchronous and asynchronous workloads are not studied. Fourth, Existing energy-conservation schemes do not take into account disk failure.

3.2 eRAID: Conserving Energy in Conventional Disk based RAID System

Amidst the aforementioned limitations of current solutions, there are several untapped or not fully developed potentials. These motivate us to develop a new energy saving policy to solve the problem.

1) With the help of inherent redundancy, it is possible to generate long idle periods for server disk to justify its high spin-up cost. By redirecting read requests to the primary disks and deferring write updates by non-volatile caching, the idle periods of mirror disks can be stretched as long as NV-RAM does not flush the dirty data.

2) The queueing theory is widely adopted to model RAID systems for performance analysis and performance prediction [10]. Queueing models can be used to quantify the performance measures of both throughput and response time. Recently researchers investigated the disk failure impact on disk array performance [11] while did not consider the third power/energy dimension. If we treat some parameters of the queuing model function as variables, e.g., the number of spinning disks in storage system, we are able to extend a performance prediction model for online control including dynamic disk power management.

Based on the aforementioned facts, we develop a complete online energy saving policy named eRAID for conventional disk based storage arrays, aiming at dynamically striking a good performance-energy tradeoff, namely, how to maximize the number of spun-down disks while satisfying constraints of multiple performance measures. The current design of eRAID adopts RAID-1 as the representative case for study and we also extended it to RAID-5.

We are among the first to exploit inherent redundancy to save energy in two RAID building blocks—RAID-1 and RAID-5 while simultaneously meeting the performance requirement. We developed a new integrated performance prediction model using a combination of queuing model and time series model and used it for online control for energy conservation. We considered disk failure in the performance prediction model as well. We developed energy models with the consideration of disk failure for both synchronous and asynchronous workloads, and evaluated our redundancy based energy saving policy on RAID-1 and RAID-5 both theoretically and experimentally. We developed a heuristic based dynamic disk power control algorithm to solve a multi-constraint optimization problem on how to select candidate disks to spin down for the purpose of saving energy. This problem can be formalized as follows:
Here, $E$, $T$, and $X$ denote energy consumption, mean response time and mean throughput. “base” represents the baseline system that employs no energy-efficiency policy. $Limit_T$ and $Limit_X$ are predefined performance degradation parameters (e.g. 10%) for mean response time and mean throughput respectively.

We evaluated our policy using trace-driven simulation. Experimental results show that eRAID can save up to 32% energy without violating predefined performance constraints. Compared with previous server side energy saving policies, eRAID has the following salient advantages: (1) it is a soft solution so that it does not require hardware updates for current storage systems or replacing all their conventional disks with multi-speed disks; (2) it is independent with host system and can be easily deployed to standard RAID systems without any change of existing disk array configurations such as data layout, or introducing any data migration overhead; (3) it strikes a good tradeoff between energy-saving and performance degradation by taking two performance metrics — average response time and throughput into consideration. Our detailed results are published in references [12] [13].

4. Conclusions

The outcomes of the research projects are deliverables of energy-efficient, high-performance storage prototype systems in our computer systems research community. A wide range of data-intensive applications as diverse as Internet computing infrastructure, scientific computing, bioinformatics and multimedia streaming in today’s multi-billion dollar industry can dramatically benefit, including storage servers, storage clusters and data centers.

5. ACKNOWLEDGMENTS

This work is supported in part by the US National Science Foundation under grants CNS-0646910, CNS-0646911/0509480 and CCF-0621526, and the US Department of Energy Early Career Principal Investigator Award DE-FG02-07ER25747.

Use footnotes sparingly (or not at all) and place them at the bottom of the column on the page on which they are referenced. Use Times 8-point type, single-spaced. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

6. References


