A Prefetching Scheme for Energy Conservation in Parallel Disk Systems

Adam Manzanares, Kiranmai Bellam, and Xiao Qin
Department of Computer Science and Software Engineering
Auburn University
Auburn, Alabama 36830, USA
{acm0008, kzb0008, xqin}@auburn.edu

Abstract

Large-scale parallel disk systems are frequently used to meet the demands of information systems requiring high storage capacities. A critical problem with these large-scale parallel disk systems is the fact that disks consume a significant amount of energy. To design economically attractive and environmentally friendly parallel disk systems, we developed two energy-aware prefetching strategies for parallel disk systems with disk buffers. First, we introduce a new buffer disk architecture that can provide significant energy savings for parallel disk systems while achieving high performance. Second, we design a prefetching approach to utilize an extra disk to accommodate prefetched data sets that are frequently accessed. Third, we develop a second prefetching strategy that makes use of an existing disk in the parallel disk system as a buffer disk. Compared with the first prefetching scheme, the second approach lowers the capacity of the parallel disk system. However, the second approach is more cost-effective and energy-efficient than the first prefetching technique. Finally, we quantitatively compare both of our prefetching approaches against two conventional strategies including a dynamic power management technique and a non-energy-aware scheme. Using empirical results we show that our novel prefetching approaches are able to reduce energy dissipation in parallel disk systems by 44% and 50% when compared against a non-energy aware approach. Similarly, our strategies are capable of conserving 22% and 30% of the energy when compared to the dynamic power management technique.

1. Introduction

The number of large-scale parallel disk systems is increasing as our need for computing power and storage space grows beyond the capacities offered by traditional servers. Typical examples of applications requiring large-scale parallel disk systems include long running simulations [4], remote sensing applications [15] and biological sequence analysis [6], to name just a few. The main drawback of applying large-scale parallel disk systems to support scientific and business applications is that as the size of parallel disk systems grow, the energy consumed by a large number of disks often becomes a large part of the total cost of ownership. Reducing the energy costs of operating these large-scale parallel disk systems often becomes one of the most important design issues for high-performance computing platforms and data clusters. Recent studies show that disk systems can account for nearly 27% of the total energy consumption in a data center [16][19]. Even worse, the push for disk systems to have larger capacities and speedier response times has driven energy consumption rates upward.

Reducing energy consumption of computing platforms has become an increasingly hot research field. Green computing has recently been targeted by government agencies and efficiency requirements have been outlined [9]. Large-scale parallel disk systems inevitably demand high-energy requirements due to the issue of scalability. Data centers typically consume anywhere between 75 W/ft^2 to 200 W/ft^2 and this may increase to 200-300 W/ft^2 in the near future (see, for example, [13]). These large-scale computing platforms not only have a large economical impact on companies and research institutes, but also produce environmental impacts. Data from the US Environmental Protection Agency indicates that generating 1 kWh of electricity in the United States results in an average of 1.55 pounds (lb) of carbon dioxide (CO2) emissions. With large-scale clusters requiring up to 40TWh of energy per year at a cost of over $4B it is easy to conclude that energy-efficient clusters can have huge economical and environmental impacts [1].

Several techniques proposed to conserve energy in storage systems include dynamic power management schemes [3][12], power-aware cache management strategies [21], software-directed power management techniques [18], redundancy techniques [14], and multi-speed settings [5][7][10]. However, the research

978-1-4244-1694-3/08/$25.00 ©2008 IEEE

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:52:50 UTC from IEEE Xplore. Restrictions apply.

on energy-efficient prefetching is still in its infancy. Therefore, it is imperative to develop new prefetching techniques to significantly reduce energy consumption in parallel disk systems while maintaining high disk I/O performance.

The objective of this study is to use the novel parallel disk architecture with buffer disks (see [19] for details of the disk architecture) to aid in the reduction of the energy consumption of large-scale parallel disk systems. To fully utilize buffer disks while aggressively putting data disks into low-power modes, we investigate an energy-aware prefetching mechanism (PRE-BUD for short) to dynamically fetch the most popular data sets from data disks into buffer disks. In doing so, it is possible to turn the data disks into the low-power modes for a longer period of time to conserve energy while having buffer disks to serve disk requests.

The PRE-BUD mechanism attempts to prefetch data into the buffer disk with the desired consequence of reducing the total energy consumption of the parallel disk system. This work aims at designing two prefetching strategies using the PRE-BUD mechanism. Specifically, the first approach using PRE-BUD adds an extra disk, which performs as a buffer disk. The second approach uses an existing disk in the system as a buffer disk. The design of these two strategies relies on the fact that in a wide variety of data-intensive applications (e.g., web applications) a small percentage of the data is frequently accessed [11]. The goal of this research is to move this small amount of frequently accessed data from data disks into buffer disks, thereby allowing data disks to switch into low-power modes.

The rest of the paper is organized as follows. Section 2 introduces and summarizes related work. Section 3 describes our pre-fetching algorithm PRE-BUD. Section 4 presents preliminary results and provides a discussion of the results. Section 5 concludes the paper by pointing out the future research directions in addition to the major contributions of this work.

2. Related Work

Recent research work has focused on energy efficient disk systems. This is fueled by the taxing energy demands of disk systems. The disregard of energy efficiency in storage systems has created a diverse research field. MAID was developed as a new approach to designing mass storage systems with a small power envelope [2]. This is a novel approach to new mass storage systems, but our solution can be readily applied to any existing large-scale parallel disk systems. Hibernator improved energy savings, while still maintaining performance [20]. This was achieved by managing multi-speed disks energy efficiently with performance metrics still in place. Our work differs from this approach in that our approaches make use of a dedicated buffer disk to accommodate popular data sets.

In recently years work has also been put into energy efficient pre-fetching schemes. For example, optimizing compilers for multi-speed disks were developed to produce energy savings [17]. Our approach is quite different from these existing approaches because ours is focused on the architecture of parallel disk systems and the management of data on all the disks in a parallel disk system. S. Huaping and M. Kuman introduced an energy efficient caching mechanism combined with a passive pre-fetching algorithm to produce energy savings for mobile distributed systems [8]. Previous studies also investigated embedded processors with energy efficient pre-fetching strategies [22]. Although the previous two studies focused on energy efficient pre-fetching, their strategies only dealt with embedded processors. In contrast, our work focuses on energy efficient pre-fetching strategies in the context of large-scale parallel disk systems supporting data-intensive applications.

More importantly, our research is novel in the sense that traditional energy saving techniques put a disk into a standby, or sleep, state only if a break-even time is met. These strategies try to extract large energy savings at the expense of reliability. Reliability of disks is directly related to power state transitions of a disk [3][12]. Our buffer disk alleviates this reliability problem by putting frequently used data into the buffer disk. The frequently used data sets are likely to come from multiple disks, but the popular data sets stored in the buffer disk make it a frequently used disk. This allows all other disks in the system to sleep longer and reduces their power state transitions, which also increases reliability of scalable parallel disk systems.

This work aims at extending our ideas of leveraging the BUD framework to conserve energy in parallel I/O systems [19]. The BUD framework allows us to use multiple disks as buffer disks for a large-scale parallel disk system. In this paper we focus only on pre-fetching in the context of the BUD framework. This will allow us to save energy and reduce state transitions for the disk system. For simplicity, we only focus on using one buffer disk for the entire disk system. This simplification allows us to demonstrate our pre-fetching approach concretely. When the single buffer disk becomes a performance bottleneck of a parallel disk system, one can either add more buffer disks into the disk system or turn several data disks into buffer disks to solve the performance bottleneck problem.

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:52:50 UTC from IEEE Xplore. Restrictions apply.

3. Energy Efficient Prefetching Strategy

The prefetching algorithm uses the frequency that a block is requested as a heuristic. The first step of the algorithm iterates over all of the requests and counts the references for each unique block. Then it sorts the list of unique blocks by the number of references to each block. At this point the algorithm puts the highly requested blocks into the buffer until it is full.

The last part of the algorithm also iterates over all requests trying to figure out how long each disk can sleep. If a block requested for a disk is in the buffer disk, the corresponding data disk can sleep longer. The buffer disk handles the request and the distance between requests on the same disk becomes cumulative. If a requested block is not in the buffer the disk must be woken up to serve the request, this is handled by the power management mechanism. The distance is then set to zero since the disk had to be woken up.

Using the frequently accessed heuristic the PRE-BUD strategy should have a small performance impact on the system. Almost all steps of the algorithm run linearly with respect to the number of requests. The only step that is not linear is the phase that sorts the list of requests according to their frequency. Sorting is a common procedure and is known to have a best-case run-time of O(nlogn), where n is the number of data request to be sorted. The PRE-BUD strategy is able to have a run time of O(n + nlogn) using an efficient sorting algorithm.

The PRE-BUD strategy is not assumed to be optimal, since the requested blocks are sorted using their frequency. The frequency is used as a heuristic to select blocks to be placed in the buffer. An optimal strategies goal would be to select the requests to be placed in the buffer that produce the largest impact on the standby time of disks. The standby time of each disk is directly related to the energy savings of the system.

4. Simulation Results

The results displayed in Figure 1 held the number of disks to 4 and also kept the data size of each request at 275MB. We have omitted the results of the non-energy aware approach, since they are constant and higher than the energy aware strategy. Buffer Strategy 1 adds an extra disk and Buffer Strategy 2 uses an existing disk as the buffer disk. As expected the performance of both of our strategies were lowered when the hit rate was decreased. Disk sleep times are lowered once a miss is encountered. This is due to the fact that a disk has to wake up to serve the request. This will increase the energy consumption of disks that have to serve the missed requests. This leads to an increase in the total energy consumption of the entire system. Our buffered large-scale parallel disk system is still able to consume less energy than the energy aware approach. The energy aware and non-energy aware disk systems are not affected by buffer disk miss rates.

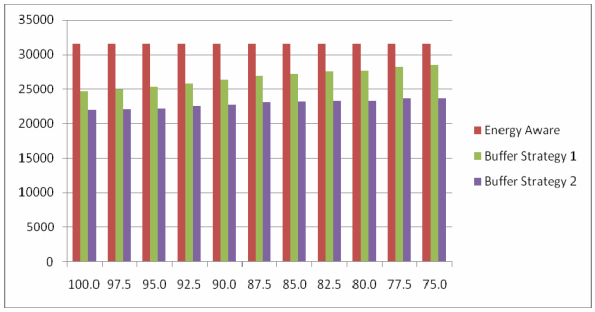


Fig. 1: Energy Savings (J) vs. Hit Rate

As the hit rate is lowered, the first buffer disk approach begins to approach the same level of performance as the energy aware strategy. It is only able to save 10% energy over the energy aware strategy when the hit rate is 75%. This is because adding the extra disk puts extra energy requirements on the system, and lowering the hit rate further impacts the energy benefits of the first strategy. The second buffered disk approach is still able perform 25% better than the energy aware approach. This is because there is not the extra energy penalty of adding an extra disk. The capacity of your disk system will be lowered using this approach.

The hit rate becomes a very important factor in the performance of our approaches. If the buffer disk is constantly missing requests then both strategies will eventually downgrade to the energy aware approach. Fortunately applications have been documented to request 20% of the data available 80% of the time. Our heuristic based approach would work considerably well in this case. This is modeled by the 80% hit rate. The buffered disk approach one and two are able to save 12% and 26% energy over the energy aware strategy when the hit rate is 80%. Similarly, they are able to save 37% and 47% of the total energy compared to the non-energy aware approach.

The first buffer disk approach downgrades more quickly than the second approach as the hit rate is decreased as compared to the energy aware approach. This is not that great of a concern since the first strategy is still able to have a positive impact on the reliability of the system as compared to the energy aware approach. The first buffer disk approach is still able to produce significant energy savings over the

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:52:50 UTC from IEEE Xplore. Restrictions apply.

non-energy aware approach without compromising the reliability of the system. The energy savings performance of the second approach does not diminish as quickly as the first approach, but there will be an impact on the capacity of the system. The second approach is also able to reduce the number of state transitions.

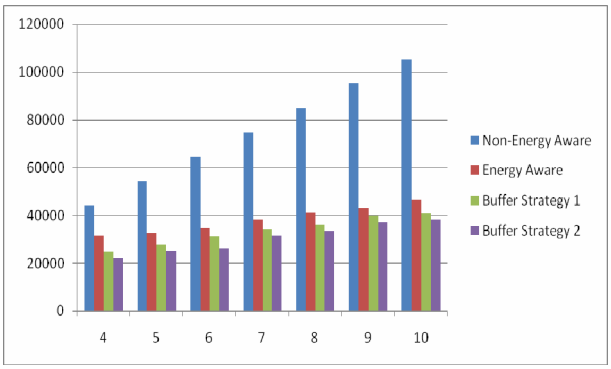


Fig. 2: Energy Savings (J) vs. Number of Disks

For the next set of experiments we increased the number of disks contained in the disk system. We kept the size of the buffer constant. This had the effect of increasing the miss rate as disks were added to the system. If a disk is added to the system it demands more blocks to be placed into the buffer. Not all of these blocks can be placed into the buffer disk. As a consequence of this, the hit rate is lowered.

From Fig. 2 we are able to see that the non-energy aware approach wastes a considerably larger amount of energy as compared to all of the energy aware approaches. This is expected since the non-energy aware approach is not able to place disks in the standby mode. Buffer strategy 1 was able to produce a 12% increase in energy savings over the energy aware strategy when 10 disks were simulated. Similarly, buffer strategy performed even better with an 18% increase. This is expected again because of the energy overhead adding an extra disk Buffer Strategy 1 requires.

Our approach produces promising results as the number of disks is increased. This is an important observation, since our target system is a large-scale parallel system. This leads us to believe our system will produce energy benefits regardless of the number of disks in a system.

5. Conclusions and Future Work

The use of large-scale parallel disk systems continues to rise as the demand for information systems with large capacities grows. Large-scale parallel disk systems combine smaller disks to achieve large capacities. The problem is that these large- scale disk systems can be extremely energy inefficient. Disk systems can account for nearly 27% of the energy demands of a large-scale computing platform. The energy consumption rates are rising as disks become faster and disk systems are scaled up. Our goal was to increase the energy efficiency of a large-scale disk system using a buffer disk that would pre-fetch frequently accessed data. This had the effect of increasing idle periods for disks facilitating long sleep times. As sleep times are increased the disk is able to save energy over being in the idle state.

We proposed two different methods of using a buffer disk. The first strategy added and extra disk to the system and the second approach used an existing disk as the buffer disk. The first strategy consumes more energy because an extra disk is added, but does not compromise the capacity of the disk system. We compared our approaches against non-energy aware and energy aware approaches. The buffered disk approaches were both able to produce energy savings as compared to the non-energy aware and energy aware strategies. The energy savings was larger when compared against the non-energy aware strategy. Although our approaches did not save as much energy as the energy aware approach, they were both able to positively impact the reliability of the disk system. Both approaches were able to lower the number of state transitions as compared to the energy aware approach.

For the future research work we would like to work on adding more than one buffer disk to the system. The number of buffer disks will have to be increased as the scale of the disk system is increased. This will add the extra requirements parallel applications demand and our strategies will have to be modified to reflect these changes. Our work also used only synthetic traces for our simulation results. We need to incorporate traces from real-world applications to improve the feasibility of our approaches. The reliability impacts of using the buffer disk can also be researched more heavily.

Acknowledgements

The work reported in this paper was supported by the US National Science Foundation under Grants No. CCF-0742187 and No. CNS-0713895, Auburn University under a startup grant, and the Intel Corporation under Grant No. 2005-04-070

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:52:50 UTC from IEEE Xplore. Restrictions apply.

References

[1] J. Chase and Ron Doyle. "Energy Management for Server Clusters," *Proc. the 8th Workshop Hot Topics in Operating Systems (HotOS-VIII*), pp. 165, May 2001.

[2] D. Colarelli and D. Grunwald. "Massive Arrays of Idle Disks for Storage Archives," *Proc. Int'l Conf. Supercomputing*, Nov. 2002.

[3] F. Douglis, P. Krishnan, and B. Marsh, "Thwarting the Power-Hunger Disk," *Proc. Winter USENIX Conf.*, pp.292-306, 1994.

[4] H. Eom and J.K. Hollingsworth. "Speed vs. accuracy in simulation for I/O-intensive applications," *Proc. Int'l Symp. Parallel and Distr. Processing Symp.*, pp. 315–322, May 2005.

[5] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Fanke, "DRPM: Dynamic Speed Control for Power Management in Server Class Disks," *Proc. Int'l Symp. of Computer Architecture*, pp. 169-179, June 2003.

[6] J. Hawkins and M. Boden, M., "The applicability of recurrent neural networks for biological sequence analysis," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 2, no. 3, pp. 243 – 253, July-Sept. 2005.

[7] D. P. Helmbold, D. D. E. Long, T. L. Sconyers, and B. Sherrod, "Adaptive Disk Spin-Down for Mobile Computers," *Mobile Networks and Applications*, Vol. 5, No.4, pp.285-297, 2000.

[8] S. Huaping, M. Kuman, S. Das, Z. Wang. "Energy-Efficient Caching and Prefetching with Data Consistency in Mobile Distributed Systems," *Proc. Int'l Parallel and Distributed Processing Symp.*, 2007.

[9] E. Jones, (2006-10-23). *EPA Announces New Computer Efficiency Requirements*. U.S. EPA. Retrieved on 2007-10-02.

[10] P. Krishnan, P. Long, J. Vitter, "Adaptive Disk Spindown Via Optimal Rent-to-buy in Probabilistic Environments," *Proc. Int'l Conf. on Machine Learning*, pp. 322-330, July 1995.

[11] T.T. Kwan, R.E. McGrath, and D.A Reed, "NCSA's World Wide Web Server: Design and Performance," *Computer*, vol. 28, no. 11, pp. 68 – 74, Nov. 1995.

[12] K. Li, R. Kumpf, P. Horton, and T. E. Anderson, "A Quantitative Analysis of Disk Drive Power Management in Portable Computers," *Proc. Winter USENIX Conf.*, pp.279-292, 1994.

[13] B. Moore. Taking the data center power and cooling challenge. *Energy User News*, August 27th, 2002.

[14] E. Pinheiro, R. Bianchini, C. Dubnicki, "Exploiting redundancy to conserve energy in storage systems," *Proc. Sigmetrics and Performance*, Saint Malo, France, June 2006.

[15] D.B. Trizna, "Microwave and HF Multi-Frequency Radars for Dual-Use Coastal Remote Sensing Applications," *Proc. MTS/IEEE OCEANS*, pp. 532 - 537, Sept. 2005.

[16] Power, Heat, and Sledgehammer. *White Paper, Maximum Institution Inc*., April. 2002.

[17] S.W. Song, M. Kandemir, "Energy Aware Pre-Fetching for Multi Speed Disks," *Proc. the 3rd Conf. Comp. Frontiers*, pp. 105-114, May 2006.

[18] S.W. Son, M. Kandemir, and A. Choudhary, "Software-directed disk power management for scientific applications," *Proc. Int'l Symp. Parallel and Distri.Processing*, April, 2005.

[19] Z. Zong, M. Briggs, N. O'Conner, X. Qin. "An Energy-Efficient Framework for Large-Scale Parallel Storage Systems," *Proc. Int'l Parallel and Distri. Processing Symp.*, March 2007.

[20] Q. Zhu, Z. Chen, L. Tan, Y. Zhor, K. Keeton, J. Wikes. "Hibernator Helping Disk Arrays Sleep Through The Winter," *ACM Symp. Operating System Principles*, October. 2005.

[21] Q. Zhu, F. M. David, C. F. Devaaraj, Z. Li, Y. Zhou, and P. Cao, "Reducing Energy Consumption of Disk Storage Using Power-Aware Cache Management," *Proc. High-Performance Computer Architecture*, 2004.

[22] X. Zhuang and S. Pande. "Power-Efficient Prefetching via Bit-Differential Offset Assignment on Embedded Processors," *Proc. ACM Conf. Languages, Compilers, and Tools for Embedded Systems,* 2004.

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:52:50 UTC from IEEE Xplore. Restrictions apply.