

Power Optimization in Disk-Based Real-Time Application Specific Systems

Inki Hong and Miodrag Potkonjak

Computer Science Department, University of California, Los Angeles, CA

ABSTRACT

While numerous power optimization techniques have been proposed at all levels of design process abstractions for electronic components, until now, power minimization in mixed mechanical-electronic subsystems, such as disks, has not been addressed. We propose a conceptually simple, but realistic power consumption model for disk drives. The core of the paper are heuristics for optimization of power consumption in several common hard real-time disk-based design systems. We show how to coordinate tasks scheduling and their disk data assignment, in order to minimize power consumption in both electronic and mechanical components of used disks. Extensive experimental results indicate significant power reduction.

1.0 Introduction

Magnetic disks are the de-facto standard for providing non-volatile high volume memory capacity in modern computer systems. Disks provide superior trade-off with respect to common design metric such as cost, memory capacity, latency, data input-output bandwidth and reliability in comparison with all other alternatives. Until recently, disks have been used mainly in general purpose computing systems. However, convergence of several application and technological trends resulted in the rapidly increasing importance of massive storage in application specific systems. There is rapid growth in applications such as internet-based servers (e.g. world wide web), video-on-demand, interactive television, and video conferencing, all of which have as dominating components large volume data management. At the same time technological trends indicate that key design metrics of modern and future application specific designs, such as speed, power, and weight, are dominated by massive storage elements. Most often, magnetic disk is already a bottleneck in current application specific computer and communication systems.

Another equally pronounced consequence of the current application and technological trends is increasing importance of power minimization. Our main strategic objective is to give impetus for research and development of synthesis and compilation techniques for design of massive storage-based application specific systems. We have three main technical goals in this paper:

1. To establish an accurate, but computationally efficient, performance and power consumption models for disk-based systems.
2. To identify most effective ways to reduce power in disk-based application specific systems.
3. To develop a practical approach and optimization synthesis algorithms for a scheduling and assignment of disk-based real-time systems.

The detailed description of the synthesis approach for optimization of disk-based application specific systems can be found in [5].

2.0 Background Material

In this section we first provide an overview of power consumption sources in a disk and briefly discuss the most popular timing models of a magnetic disk. We conclude the section, by explaining the selected hardware and computational models. The detailed description of disk technology is available in [5].

Power required by a hard disk drive is consumed by its many different components. To complicate matters even further, the power requirements of each component will vary with the current operational mode of the disk. Common operational modes with different power requirements are: Start-up, Seek, Read/Write: Idle, Standby and Sleep. In each of the distinct operational modes available, a different amount of strain is placed upon each of the individual disk components, varying the amount of power consumed [5]. A seek moves the disk

seek distance range [tracks]	seek time
1 - 50	$1.9 + \sqrt{\text{distance} - \text{distance}/50}$
51 - 100	$8.1 + 0.044 * (\text{distance} - 50)$
101 - 500	$10.3 + 0.025 * (\text{distance} - 100)$
501 - 884	$20.4 + 0.017 * (\text{distance} - 500)$

Table 1 Typical seek time for an IBM disk.

head (arm) from track to track. Several techniques have been proposed for analytic and empirical modeling of access data [13, 18]. The common denominator in all of them is that longer distance which arm has to travel corresponds to larger time overhead. Typical seek times for an IBM disk are given in Table 1 [11].

Our selection process of computational and hardware models was mainly guided by the goal to cover as large as possible set of modern and future disk-based application specific systems. The system has three components: disk, main memory, and processor. Processor by itself can have multiple processors and/or ASICs. Since for power minimization in both memory and processor (and ASIC) several approaches are readily available [12], we focus our attention on disk's power optimization. We assume that a disk is a separate unit, as it is almost always the case in industrial practice.

We assume that each of tasks follows homogeneous synchronous data flow semantics and syntax individually [9]. We assume, with no loss of generality that all tasks have identical periods. When this is not the case, a simple preprocessing step and application of the least common multiple (LCM) theorem [8], in polynomial time transforms an arbitrary set of periods to this design scenario. We assume no task preemption. Note that this preemption restriction, actually does not impact any of the proposed methods, since in all discussed design cases non-preemptive policies yield superior results in comparison with preemptive policies.

Furthermore, each task has a need to read or read and write data to the disk. We assume that for each task a sequence of disk blocks to be accessed for its execution is given. Time to serve one read or write request is the sum of seek time and data transfer time. Seek time is proportional to distance which disk's head has to travel, and read and write time is proportional to amount of data which has to be transferred. The typical seek times for an IBM disk in Table 1 have been used for experiments. The goal is to properly schedule all tasks and their required data transfers, so that all timing constraints are satisfied and disk's power consumption is minimized.

3.0 Related Work

Although there have been constant stream of proposed alternative massive storage technologies, magnetic disks have dominated secondary storage since the mid sixties. Detailed description of magnetic disks can be found in many books [6]. Another brief, but excellent exposition are papers [13] and [4]. An introductory exposition of basic disk principles is also given in modern architecture and operating systems textbooks [11]. Wood and Hodges [17] survey state-of-the-art and technology trends in direct access storage devices, mainly magnetic disks. Disk modeling recently attracted a great deal of interests [4,11,13,18].

The early disk-related research in operating systems has been focused on development of scheduling algorithms for efficient use of high-volume storage in time-shared mainframes [16]. Later, operating systems

researchers developed new disk scheduling algorithms for new general-purpose computing platforms assuming increasingly more realistic and complex disk models [14]. Grossman and Silverman discussed placement of records on a secondary storage device to minimize access time [3].

Recently a number of synthesis and compilation techniques for power optimization at all levels of abstractions during design process have been proposed [1, 15]. Although, power optimization is most effective at the higher levels of abstractions, until recently majority of power minimization techniques were proposed at logic synthesis and physical design phases of design [1, 12, 15]. A good survey of low power storage alternatives for general purpose mobile computing is given in [2].

4.0 Disk Power Model

Our model separately considers two subparts: mechanical and electronic subsystems. Those two parts have two sharply different power dependencies [4,5].

The electronic part follows standard power trade-offs of CMOS-based designs. The sources of power consumption in a CMOS integrated circuit are due to four types of currents: leakage, standby, short-circuit, and capacitive. All the currents except capacitive can be reduced to a relatively low percentage of the total design power by a combination of proper design techniques [12, 15] and are mainly independent from the synthesis tasks related to architectural and application design of disks electronic subsystems. Therefore, the power consumption can be quantified using the following widely quoted equation:

$P_{elect} = \alpha * C * V_{dd}^2 * f$ where α is the activity factor, C is average capacitance switched per cycle, V_{dd} is the supply voltage, and f is the cycle frequency, assuming that V_{sw} the switched voltage is equal to the supply voltage. For power-delay dependency, we use the 6th order Nevine's rational polynomial approximation proposed and experimentally verified by Chandrakasan et al. [1].

Elaborate measurements [4] show linear dependency between rotational spindle motor speed and power. In particular, we use the following formula, derived from [4]:

$P_{disk} = P_{fs} - \gamma * (nrs - ors)$, where P_{disk} is power consumption of the disk which operates on operating rotations speed, denoted ors , γ is constant scaling coefficient, P_{fs} is power consumption at nominal rotational operating speed, denoted by nrs .

We selected parameters in this formula, to follow our conservative estimation of improvements in power consumption. We used the following values in our experimentations: $P_{fs} = 700 \text{ mW}$; $\gamma = 110 \text{ mW}/1000 \text{ rpm}$; and $nrs = 5000 \text{ rpm}$.

5.0 Optimization: Approach, Problem Formulation, and Optimization Strategy

We now summarize our approach to power minimization. The key idea is to minimize seek time using proper scheduling and data assignment algorithms so that disk read/write time can be slowed down to result in the opportunities of exploiting power optimization degrees of freedom; the voltage of the electronic components can be reduced and the spindle motor speed of the mechanical component can be slowed down.

The most general version of the targeted problem can be formulated in the following way:

Problem: The Power Optimization Under Throughput Requirement Using Disk Seek Time Minimization, Spindle Motor Speed Scaling and Supply Voltage Scaling.

Instance: Given a set of M tasks described by the disk block access sequence, an initial voltage V , an initial spindle motor speed S and positive constants D and P .

Question: Are there a disk block assignment, a static

periodic schedule of the tasks, a new spindle motor speed S' and a new voltage V' such that the disk seek time + read/write time is at most D and the power consumption is at most P ?

We proved that our problem is NP-complete [5]. We solve the power optimization problems in two steps. First, we find a task schedule and a disk assignment such that the disk seek time is minimized. Next, a voltage scaling and a spindle motor speed scaling are performed such that the throughput requirement is met.

Since the computational complexity of the disk head movement minimization problem forbids an exact or optimal solution, effective heuristic methods have been developed for the problem. The task scheduling problem is transformed into a TSP problem and an efficient and effective TSP heuristic [10] is applied to the transformed problem. For the disk assignment problem, the simulated annealing (SA) algorithm [7] has been used. The detailed description of the TSP and SA heuristics uses is given in [5]. The task scheduling and disk assignment problem

Number of Tasks	Task Scheduling Problem				Disk Assignment Problem				Task Scheduling and Disk Assignment Problem			
	Random		Optimized		Random		Optimized		Random		Optimized	
	Average	Best	Average	Best	Average	Best	Average	Best	Average	Best	Average	Best
50	529.69	521.16	355.81	355.81	543.24	516.57	420.54	419.44	539.73	521.49	311.83	308.80
100	1345.15	1328.51	865.25	865.25	1355.70	1308.31	980.69	977.22	1367.57	1330.07	693.07	679.21
150	2188.40	2133.48	1384.35	1384.29	2170.07	2112.88	1481.71	1465.89	2171.83	2116.32	1004.37	985.16
200	3165.81	3125.58	1896.31	1895.76	3217.62	3141.43	2136.19	2112.71	3242.36	3189.99	1373.38	1358.98
250	3877.75	3850.98	2182.20	2181.90	3934.39	3866.63	2488.37	2429.78	3944.79	3902.58	1555.03	1536.65

Table 2 The results for the disk seek time and read/write time minimization.

Number of Task	Task Scheduling Problem	Disk Assignment Problem	Task Scheduling and Disk Assignment Problem
50	0.65	48.04	19.26
100	2.96	144.12	50.68
150	9.10	293.90	102.27
200	9.59	578.42	198.13
250	16.03	1063.90	388.49

Table 3 Running Time for example from Table 2 (seconds on SUN SPARCstation 4)

Number of Task	Task Scheduling Problem			Disk Assignment Problem			Task Scheduling and Disk Assignment Problem		
	Optimized PD			Optimized PD			Optimized PD		
	Electronic	Spindle	Total	Electronic	Spindle	Total	Electronic	Spindle	Total
50	227.54	409.77	637.31	323.20	482.07	805.27	187.38	375.68	563.06
100	185.33	373.87	559.20	239.67	419.45	659.12	143.03	330.52	473.55
150	169.42	358.67	528.09	189.94	377.97	567.91	125.27	308.51	433.78
200	148.09	336.31	484.40	169.65	358.90	528.55	112.67	290.80	403.47
250	130.87	315.75	446.62	145.73	333.60	479.33	104.58	278.32	382.90

Table 4 The results for the power minimization using voltage scaling and spindle motor speed scaling.

employs a reiterative heuristic which repeatedly solves the task scheduling problem and the disk assignment problem separately using their TSP and SA heuristics until no improvement is achieved. The heuristic is described using the following pseudo-code:

```
Generate a random disk assignment.
Apply the TSP heuristic to find a task
schedule given the random disk assignment.
Set the current schedule and assignment to
the best-so-far solution.
Repeat
  Apply the SA algorithm to find a disk
  assignment given the current task schedule.
  If the new assignment does not improve upon
  the best-so-far, stops the loop and return
  the best-so-far.
  Apply the TSP heuristic to find a task
  schedule given the current disk assignment.
  If the new schedule does not improve upon
  the best-so-far, stops the loop and return
  the best-so-far.
```

6.0 Experimental Results

We have generated random examples by varying the number of tasks. The number of blocks and the schedule period are chosen to be the same as the number of tasks. We have tried the examples of 50, 100, 150, 200, and 250 tasks. Each task accesses either one or two blocks. Each disk block access involves a disk read and write. The disk seek time + read/write time of the best random solution is used as a deadline. The initial power dissipation (PD), the initial PD by the spindle system, the initial PD by the electronic part, the initial supply voltage, the initial spindle motor speed and the initial disk read/write time has been set to 1.51 W, 700 mW, 810 mW, 5.0 V, 5,000 RPM and 1.0 ms, respectively. The Tables 2 and 4 illustrate the effectiveness of the power optimization using disk seek time minimization, voltage scaling, and spindle motor speed scaling. The power consumption reductions by factors of 3.13, 2.86, and 3.66 are achieved for the task scheduling problem, the disk assignment problem and the task scheduling and disk assignment problem, respectively. Table 3 illustrates the efficiency of the proposed heuristics and running times are on SUN SPARCstation 4 with 32 MB of main memory. Even on this relatively modest platform, the large instances of the problem has been solved in relatively short run-times.

7.0 Conclusion

We studied a new problem of power optimization in disk-based application specific systems. We proposed a conceptually simple, but realistic power consumption model for disk drives. Simulated annealing and traveling salesman problem heuristics are used as optimization mechanisms for power minimization in several common hard real-time disk-based systems design scenarios. We demonstrated how to coordinate tasks scheduling and their disk data pattern access and assignment, so to minimize

power consumption in both electronic and mechanical components of used disks. Extensive experimental results indicate significant power reduction ability of the proposed techniques and algorithms.

8.0 References

- [1] A.P. Chandrakasan, et al. "Optimizing Power Using Transformations", IEEE Transactions on CAD, Vol. 14, No. 1, pp. 13-32, January 1995.
- [2] F. Douglass, et al. "Storage alternatives for mobile computers", USENIX Symposium on Operating Systems Design and Implementation (OSDI), pp. 25-37, 1994.
- [3] D. D. Grossman, H. F. Silverman, "Placement of records on a secondary storage device to minimize access time", Journal of the ACM, Vol. 20, No. 3, pp. 429-438, 1973.
- [4] E.P. Harris, et al. "Technology Directions for Portable Computers", Proc. of the IEEE, Vol. 83, No. 4, pp. 636-658, 1995.
- [5] I. Hong, M. Potkonjak, "Power Optimization in Disk-Based Real-Time Application Specific Systems", UCLA, CS Dept. Technical Report 960025, 1996.
- [6] F. Jorgensen. "The complete handbook of magnetic recording", TAB Books, New York, NY, 1996.
- [7] S. Kirkpatrick, C. Gelatt, M. Vecchi, "Optimization by Simulated Annealing", Science, Vol. 220, No. 4598, pp. 671-680, 1983.
- [8] E. L. Lawler, C.U. Martel, "Scheduling periodically occurring tasks on multiple processors", Information Processing Letters, Vol. 12, No. 1, pp. 9-12, 1981.
- [9] E.A. Lee, T.M. Parks, "Dataflow Process Networks", Proc. of the IEEE, Vol. 83, No. 5, pp. 773-799, 1995.
- [10] O. Martin, S. W. Otto, E. W. Felten, "Large-Step Markov Chains for the TSP Incorporating Local Search Heuristics", Operations Research Letters, Vol. 11, No. 4, pp. 219-224, 1992.
- [11] D.A. Patterson, J.L. Hennessy, "Computer Architecture: A Quantitative Approach", Morgan Kaufmann, San Mateo, CA, 1990.
- [12] J. Rabaey, M. Pedram, ed., "Low power design methodologies". Kluwer, Boston, MA, 1995.
- [13] C. Ruemmler, J. Wilkes, "An introduction to disk drive modeling", IEEE Computer Magazine, Vol. 27, No. 3, pp. 17-28, 1994.
- [14] M. Seltzer, P. Chen, J. Ousterhout, "Disk Scheduling Revisited", Proc. of USENIX, pp. 313-323, 1990.
- [15] D. Singh et al., "Power Conscious CAD Tools and Methodologies", Proc. of the IEEE, Vol. 83, No. 4, pp. 570-594, 1995.
- [16] T.J. Teorey, T.B. Tinkerton, A comparative Analysis of Disk Scheduling Policies, Communications of the ACM, Vol. 15, No. 3, pp. 177-184, 1972.
- [17] C. Woods, P. Hodges, "DASD Trends: Cost, Performance, and Form Factor", Proc. of the IEEE, Vol. 81, No. 4, pp. 573-585, 1993.
- [18] B.L. Worthington, G.R. Ganger, Y.N. Patt, J. Wilkes, "Online extraction of SCSI disk drive parameters", Performance Evaluation Review, Vol.23, No. 1, pp.:146-56, 1995.