A Methodology for Generating Disk Drive Energy Models Using Performance Data

Anthony Hylick and Ripduman Sohan
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge, UK CB3 0FD
firstname.lastname@cl.cam.ac.uk

ABSTRACT

The ability to make intelligent decisions with respect to reducing the energy consumption of mechanical disk drives is dependent upon accurately tracking and accounting the runtime energy consumption of the mechanical and electrical components of the drive in both active and idle states. This work outlines a simple methodology for creating accurate hard drive runtime energy models through the use of easily obtainable data derived from published specifications and performance measurements. We use the generated models to create a TRADE estimator (TRAce-Driven disk drive Energy estimator). Validation of our estimator against a diverse range of disk drives shows that results obtained from the models are within 5% of measured results.

Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance Of Systems

General Terms

Algorithms, Design, Experimentation, Measurement

Keywords

Hard Drive, Power Measurement, Energy, Model

1. INTRODUCTION

Reducing the energy footprint of computation continues to be a major topic of interest for both research and industry [3]. However, our ability to make accurate decisions with respect to reducing the energy footprint of our IT infrastructure is hampered by our poor understanding of the runtime energy consumption of the infrastructure.

A fundamental reason for our limited understanding of runtime energy consumption is the fact that we do not have complete and accurate information on how hardware devices consume energy in relation to usage and workload. There are various reasons for this knowledge gap—the required information may be difficult to accurately measure or account, require expensive and complicated runtime support, or, for the case of legacy devices, be unavailable as it was not considered necessary when the device was designed and manufactured.

We feel that having detailed and accurate knowledge of the energy consumption of mechanical hard drives would enable the ability to accurately account for runtime energy consumption. Comprehensive, runtime energy consumption data would allow for a higher degree of confidence in the energy-efficiency of storage and workload configurations before the energy bill arrives. In our previous work, we built specific hardware to enable measurement of the runtime power consumption of hard drives at a fine-grained level [4]. Our goal in this paper is to show how the intrinsic energy consumption metrics of *any* drive can be derived by characterizing its bandwidth and seek performance profiles from readily accessible information and combining them with published power consumption information. We further construct our TRADE estimator (**TRA**ce-**D**riven disk drive **E**nergy estimator), an accurate, runtime drive energy model based on state and internal operation policy (i.e. disk head location and on-disk cache maintenance). Finally, we validate our runtime model against measured results proving that the model matches with less than 5% error, enabling our primary intention of providing a simple and accurate mechanism for modeling the energy consumption of hard drives without specialized instrumentation measuring power.

Our approach most closely resembles the approach taken with the Dempsey project [9] which outlined a technique for modeling the energy consumption of hard drives using measured power characteristics. However, we differ from the Dempsey project in three principles: (*i*) Our approach does not require obtaining physical measurements of device power consumption, (*ii*) our approach does not require understanding the physical disk layout and (*iii*) our model is significantly simpler in terms of implementation and computational overhead.

HotPower '09, October 10, 2009, Big Sky, Montana, USA

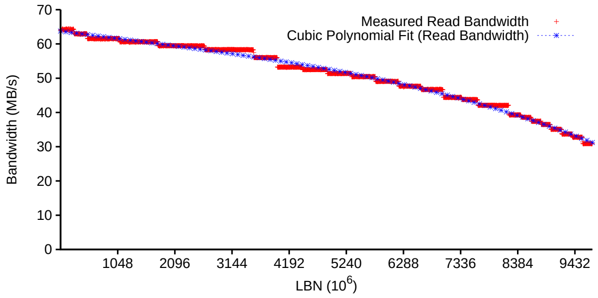

Figure 1: Measured Read Bandwidth

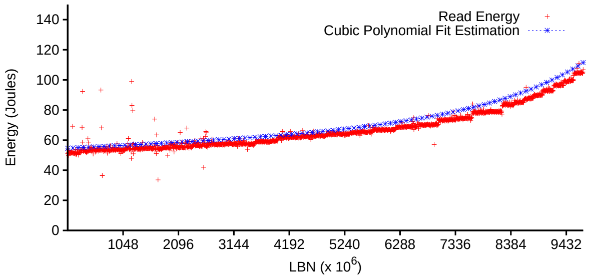

Figure 2: Estimated and Measured Read Energy

2. RELATING PERFORMANCE TO ENERGY

While the physical architecture of mechanical drives is composed of tracks on platters assembled into a spindle, it has been established that it is more efficient for drives to export an abstracted interface for simplicity and flexibility [8]. For this reason, all modern hard drives export a linear array of *logical block numbers* (LBNs) as the primitive for data storage and access.

For the purposes of efficiency, virtually all hard drives map LBNs from the outer tracks inward. Thus, lower LBNs correspond to the physical locations farthest from the spindle while higher LBNs are mapped closer to the spindle. Where multiple platters are present, drive manufacturers map sequential LBNs on tracks on multiple platters before moving inward, preserving the property of mapping lower LBNs on outer tracks.

This radial geometry yields two interesting performance characteristics: (*i*) bandwidth decreases with increasing LBN and (*ii*) seek time between two LBNs is a function of the number of cylinders spanned seeking from the originating LBN to the destination LBN. More succinctly, bandwidth decreases as LBN increases and seek time increases with the number of LBNs spanned.

In our previous work, we measured several different drives and found that while the absolute figures involved in bandwidth variation and seek time are dependent on a variety of factors (e.g. the recording density and rotation speed), the general behavior regarding the noted performance characteristics is accordant [4]. We exploit these characteristics as the basis for fingerprinting hard drive energy consumption as described below.

2.1 Estimating Transfer Energy

The energy required for transferring data to and from the disk drive is a fundamental component of overall energy consumption. It is our observation that the transfer energy can be derived from the drive bandwidth profile. Figure 1 illustrates the bandwidth profile of our sample drive[1]. This bandwidth profile was generated using raw I/O transfers (256 MB in size) across the entire disk. The figure shows bandwidth decreasing as LBN increases. The figure also shows that bandwidth remains constant for series of continuous LBNs before decreasing, forming a series of plateau decreases in bandwidth.

The plateaus are the result of *zoned bit recording*, a technique that employs storing more data on outer cylinders to increase data storage efficiency. In this method, the drive is divided into *zones*, each being composed of a number of cylinders. All the cylinders in a zone contain the same number of sectors (logical blocks), and every zone has a unique number of sectors per cylinder [2]. As a result of this recording technique, the bandwidth derived from LBNs situated in cylinders in a particular zone is constant. Note that the number of cylinders per zone and number of zones per drive is manufacturer (and drive) dependent. This characteristic makes it difficult to predict bandwidth at any given point on the drive without prior measurement.

However, upon further investigation during this work, we have seen bandwidth may be approximated by modeling it as a cubic polynomial function, the coefficients of which are easily obtained through standard curve fitting techniques. Figure 1 illustrates the use of this technique to approximate the bandwidth profile of our sample drive. We have tested this fitting strategy across ten varied drives and found that using a cubic polynomial produces the lowest root mean squared error values. This cubic approximation enables us to estimate the transfer bandwidth for any set of LBNs read or written on the disk. We are currently working on understanding and explaining this cubic relationship exactly, but we feel the combination of decreasing circumference, decreasing linear velocity, and zoned bit recording are primary contributors.

Knowing the approximate bandwidth at a given LBN, it is possible to calculate energy consumption for data transfers beginning at that LBN as a function of the *time*

[1] Throughout this paper we use measurements obtained from the Hitachi Deskstar E7K500

spent transferring the data. Energy is calculated as:

$$E_{Active} = \frac{S}{B} \times P_{Active}, \tag{1}$$

where E_{Active} is active energy, in Joules; S is filesize, in MB; B is bandwidth, in MB/s; and P_{Active} is the active power, in Watts, as provided by the drive manufacturer. Note that while the active power is usually quoted as a constant figure, our previous work showed that it actually varies according to LBN [4]. However, this variation is quite small (usually less than 2 Watts across all LBNs) and may be ignored with little loss of accuracy.

With the active power, transfer energy is calculated as an inverse of transfer bandwidth using Equation 1. Due to the accurate bandwidth model and the small variation in active power, we have found that a transfer energy model obtained in this manner provides useful estimates as Figure 2 indicates, using an example derived from the parameters obtained from our sample drive. The energy measurements in Figure 2 were taken while the bandwidth profile (Figure 1) was being generated.

2.2 Estimating Seek Energy

As outlined previously, the mapping employed by modern hard drives results in seek time being a function of the cylinders spanned seeking between a given pair of LBNs. Previous work has established that seek time increases with the square root of the number of cylinders spanned plus the head settle time [7]. Maximum head seek velocity is a function of the maximum power consumed by the actuator motor, which is in turn limited by the motor design [7]. Consequently, seek time increases with the seek distance. Our model uses seek time as an estimator for seek energy. Thus, seek time is calculated as:

$$T_{Seek} = a\sqrt{abs\left(\left\lfloor \frac{LBN_{dest.}}{SPT} \right\rfloor - \left\lfloor \frac{LBN_{start}}{SPT} \right\rfloor\right)} + s, \tag{2}$$

where T_{Seek} is seek time, in milliseconds; a is a constant coefficient; $LBN_{dest.}$ is the destination LBN; LBN_{start} is the start LBN; SPT is the number of sectors per track; and s is the head settle time, in milliseconds. The floor division represented in the equation is our way of estimating the cylinder of an LBN.

The absolute value of the difference between the destination cylinder and the start cylinder from Equation 2 provides the cylinder distance needed to approximate seek time as described in [7]. The absolute value is taken understanding that the distance will be either positive or negative, indicating a seek inward (closer to the spindle) or outward, respectively. The number of sectors per track is a figure listed in modern drive datasheets.

The constant coefficient, a, is dependent on the drive and may be obtained using the single-track and full-stroke seek times with an equation of the form:

$$a = \frac{(z - t)}{\sqrt{d_{Max}} - 1}, \tag{3}$$

where t is the single-track seek time, in milliseconds; z is the full-stroke seek time, in milliseconds; and d_{Max} is the maximum number of cylinders that can be spanned (i.e. the number of cylinders the drive has). The single-track and full-stroke seek times, as well as d_{Max}, are published numbers that can be found in the datasheet.

The final value, head settle time (s), can be obtained by plugging in the previously determined values into the equation below.

$$s = t - a, \tag{4}$$

where s is the head settle time, in milliseconds; t is the single-track seek time, in milliseconds; and a is the coefficient obtained from Equation 3. The head settle time, s, becomes more significant for short seeks as it dominates the seek time [7].

We observed that the majority of drive manufacturers provide seek power as standard information available in the datasheet, and we have seen it to be independent of LBN and seek distance in our past [4] and current measurements. Thus, provided the evidence above, our estimate of seek energy is reduced to the following:

$$E_{Seek} = T_{Seek} \times P_{Seek}, \tag{5}$$

where E_{Seek} is seek energy, in Joules; T_{Seek} is the seek time, in seconds; and P_{Seek} is the seek power, in Watts. Substituting appropriately for T_{Seek}, using Equation 2, it is possible to accurately predict the energy required to seek between any two LBNs on the disk.

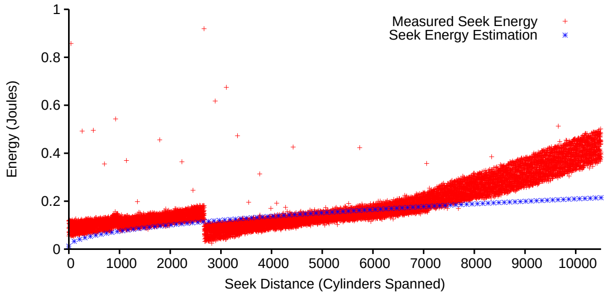

Figure 3: Seek Energy Approximation

Figure 3 provides a comparison of the measured and estimated seek energies across our sample drive. What looks to be a linear divergence beginning around 7000 cylinders is the result of the additional power needed to maintain the disk heads at higher LBNs while servicing a request, and not additional power for seeking. We encountered difficulty in taking accurate and exclusive seek measurements due to the fact that the ATA specification does not provide a mechanism to issue seeks without a request. Because of the increase in head- positioning power and decreased bandwidth, the seek measurements in Figure 3 at higher LBNs include more energy that represents non-seek activity which is difficult to accurately subtract. However, the maximum difference between the measured and estimated results shown in Figure 3 is still less than 1 Joule.

3. TRADE ESTIMATOR

While the previous section outlined the methodology for obtaining the drive transfer and seek energy profiles, this section outlines the issues associated with modeling runtime energy consumption.

The total amount of energy consumed by a drive servicing a set of N requests (including I time idle) comprised of S seeks may be modeled as follows:

$$E_{Total} = \sum_{i=0}^{N} E_{Active} + \sum_{i=0}^{S} E_{Seek} + \sum_{i=0}^{I} E_{Idle}, \quad (6)$$

where E_{Total} is the total energy, in Joules; E_{Active} is the active energy, in Joules; E_{Seek} is the seek energy, in Joules; and E_{Idle} is the idle energy, in Joules. We detail the salient issues associated with every component below.

3.1 Idle Energy

Idle energy refers to the energy consumed while the disk is in an *on* state but not servicing any requests. An idle drive will consume a certain (drive specific) amount of power for the purpose of maintaining a ready state. The energy consumed by this operation is unavoidable and usually distributed in keeping the spindle spinning and the controller electronics powered.

Idle energy consumption may be considered the baseline energy consumption for the drive at all times. Any energy consumed transferring data or seeking is considered *additional* to idle energy consumption. As the idle power consumption is an important indicator of drive power efficiency, all drive manufacturers provide this information in the published drive datasheet. Thus, the total idle energy for a given time period is easily calculated by obtaining the product of the time the drive is powered and the idle power consumption.

3.2 Seek Energy

In the simplest case, calculating seek energy is driven by accurate tracking of the disk head location for the purpose of determining seek distance. However, calculating seek energy may be complicated by the possibility of requests spanning multiple tracks. Due to the black-box nature of our approach to modeling energy consumption, it is difficult to determine when track or head switches occur on transferring a series of sequential LBNs, meaning that we are unable to account for the energy associated with these activities. However, as highlighted in Section 2.2, seek energy is proportional to the number of cylinders traversed and, therefore, on transferring a sequential set of LBNs, the seek energy of a track switch is insignificant compared to transfer energy because the seek energy for close LBNs is very low. We also have not seen any significant increase in power during head switches as this simply requires activating a different head on another platter surface.

It has previously been noted [1] (and our instrumentation has confirmed) that the actuator only consumes additional power when seeking from lower LBNs to higher ones. Conversely, seeking from higher to lower LBNs does not require additional power because the spring-loaded design of the actuator ensures the return of the disk heads to their natural position on the outer edges of the platters. Hence, it is only necessary to account for seek energy when seeking from lower to higher LBNs.

3.3 On-disk Cache & Request Reordering

Modern hard drives are equipped with several megabytes of on-disk cache. This cache can significantly affect the energy consumption of the drive, and it is important that the cache is accurately accounted and modeled. For our sample drive, we found that a 512KB linear segment would be read ahead into the on-disk cache for any read request not serviced from the on-disk cache. System buffer caches can be much larger than on-disk caches and significantly reduce the chances of on-disk cache hits [5], but our objective to model drive energy consumption completely required this information.

We observed that the energy required to service requests from the cache is negligible, requiring runtime energy additions to properly reflect a full or partial cache hit or a transfer from the platters. Conversely, there is no advantage for transferring data to the cache as all data written to the cache will eventually be committed to the platters and will incur transfer energy overhead.

The behavior of the on-disk cache is proprietary and may not be publicly available. Previous work has highlighted techniques for fingerprinting on-disk caches [6]. We have successfully used these techniques to experimentally determine the size, behavior, and semantics of the on-disk cache.

Similarly, the internal request reordering employed by SCSI and SATA drives may cause the predicted energy consumption to deviate from the actual runtime energy consumption. While we have yet to accurately instrument the effect of internal request reordering, we note that if the reordering algorithm is unpublished, it is likely that, for the purposes of efficiency, the drive will strive to minimize seeks. Thus, given a set of requests, optimizing to minimize seeks is likely to match the internal reordering as carried out by the drive. Further work is required to verify this hypothesis.

4. VALIDATION

We sought to validate our model using a black-box approach by selecting three drives, picked at random, over a wide range of age and capacity properties. After instrumenting the transfer bandwidth, we estimated the transfer and seek energy consumption profiles using the method outlined in Section 2. We then built models for the drives as outlined in Section 3 and emulated the replaying of a known workload. We compared this result to drive energy consumption measurements using the equipment described in our previous work [4].

Our validation workload is composed equally of reads and writes over the entire capacity of the drive. A total of 5GB is transferred with request sizes varying between 512 bytes and 1MB. The workload is designed to cause seeks while spanning the entire addressable LBN range exported by the drives. Table 1 provides the specifics for the drives tested.

Table 1: Details of Drives Tested

Make & Model	Capacity (GB)
Seagate ST380215A	80
Hitachi Deskstar E7K500	250
Samsung HD501LJ	500

Validation was carried out by comparing the emulation energy estimates for the complete workload against those obtained by our measuring equipment. Table 2 presents the results of our TRADE estimator compared to the measured results *and* the results obtained by using the typical naïve approach of simply multiplying the trace duration by the active power.

The results in Table 2 were obtained by running ten different workload traces on each drive while capturing the power consumption and logging the trace for later replay in our emulator. The numbers shown are the average percent errors of the model energy estimates (and the naïve energy estimates) compared to the measured energy value over the ten runs. We also present the standard deviations for the error percentages to show the consistency of estimations.

Our TRADE estimates were more accurate than the naïve approach for all workloads. The accuracy of our TRADE estimator relies upon accurate power numbers in manufacturer datasheets to provide accurate energy estimates, and so, the variation in accuracy across drives seen in our results is indicative of how closely datasheet power figures match measured numbers. In the worst case, all three drive's estimated results were within 5% of the measured values compared to 12% for the naïve approach. In essence, the results show that using our model provides a more accurate estimate of the drive's runtime energy consumption.

Table 2: Comparison of Results

Drive	TRADE	std.	Naïve	std.
	Error (%)	dev.	Error (%)	dev.
Seagate	3.42	1.63	10.59	1.73
Hitachi	-0.81	0.48	9.17	0.44
Samsung	-0.57	0.80	-5.49	0.59

5. CONCLUSION

This work has detailed the possibility and accuracy of generating energy consumption estimates from performance characteristic models. These models are useful in the real world where instrumenting hard drives for energy information is not possible. Our approach saves valuable time and effort in providing energy estimates that are reliable enough upon which to make decisions.

Acknowledgements

This work would not have been possible without the help and support of Andy Hopper. Thanks to Robert Harle for providing essential feedback and comments.

6. REFERENCES

[1] Storage Review: Head Actuator. `http://www.storagereview.com/guide2000/ref/hdd/op/actActuator.html`. Retrieved 28th July 2009.

[2] J. L. Hennessy and D. A. Patterson. *Computer architecture: a quantitative approach*. Morgan Kaufmann, 3rd edition, 2002.

[3] A. Hopper and A. Rice. Computing for the future of the planet. *Transactions of the Royal Society*, 366(1881), 2008.

[4] A. Hylick, R. Sohan, A. Rice, and B. Jones. An analysis of hard drive energy consumption. MASCOTS, 2008.

[5] J. Park and K. Koh. A space-efficient on-disk prefetching algorithm. ICCSA, 2007.

[6] F. I. Popovici, A. C. Arpaci-Dusseau, and R. H. Arpaci-Dusseau. Robust, portable I/O scheduling with the disk mimic. USENIX, 2003.

[7] C. Ruemmler and J. Wilkes. An introduction to disk drive modeling. *IEEE Computer*, 27(3), March 1994.

[8] A. Silberschatz, P. B. Galvin, and G. Gagne. *Operating System Concepts*. Addison Wesley, 8th edition, 2008.

[9] J. Zedlewski, S. Sobti, N. Garg, F. Zheng, A. Krishnamurthy, and R. Wang. Modeling hard-disk power consumption. FAST, 2003.