2014 IEEE 22nd International Symposium on Modelling, Analysis & Simulation of Computer and Telecommunication Systems

Which Storage Device is the Greenest? Modeling the Energy Cost of I/O Workloads

Yan Li, Darrell D. E. Long
Storage Systems Research Center, University of California, Santa Cruz
{yanli,darrell}@cs.ucsc.edu

Abstract—**The performance requirements and amount of work of an I/O workload affect the number of storage devices and the run time needed by the workload, and should be included in the calculation of the cost or energy consumption of storage devices. This paper introduces models to calculate the cost and energy consumption of storage devices for running a variety of workloads, categorized by their dominant requirements. Measurements of two latest hard disk and solid-state drive (SSD) are included to illustrate the models in practice. Contrary to common belief, SSD is not the energy efficient choice for many workloads.**

Keywords*—*energy measurement; storage system; measurement; measurement techniques

I. Introduction

Energy consumption comprises a large fraction of the running cost of large-scale storage systems [12]. Estimating and calculating the energy consumption of storage devices (device energy use) are important tasks for system designers and administrators. Existing studies of device energy use mainly focused on the power of the device without paying enough attention to the amount of work required by the workload. Real-world I/O workloads always include a certain amount of work. The power of the storage devices alone cannot reflect the energy consumed to do these work, because the device energy use for running an I/O workload (E) equals the integral of a device's power (P) times the number of devices (N) over the duration (T):

$$E = \int_{T} PN \tag{1}$$

It is meaningless to say that, for example, "solid-state drives (SSD) are more energy-efficient than hard disk drives (HDD)" without considering the amount of work in the I/O workloads; SSDs are generally lower in power, but they are often lower in storage density, and sometimes multiple SSDs are needed to match the storage capacity of one HDD.

In this paper, we construct models to calculate the device energy use for doing a certain amount of I/O work. First, I/O workloads are categorized into capability or capacity workloads by their most demanding requirements. The capability workloads require the storage system to provide a specific amount of I/O bandwidth/IOPS/latency. Typical web/file/database server workloads are such capability workloads because they need to sustain a specific amount of bandwidth. The capacity workloads require the storage system to provide a specific amount of storage capacity. Most workloads from archival storage systems and backup systems belong to this category.

Models are then constructed for each category of workloads. Given the specific requirements of a workload, we calculate the number of devices needed, the run time needed, and what power management scheme should be applied. The models also include key performance and power properties of the storage devices that need to be measured from the devices. The results from the models can be used to find the most energy efficient or cost efficient device for a specific workload. Measurements of two SSD and HDD devices are included to illustrate the models in practice. We also measure the energy consumption of several HDDs' power-cycle and show that the results are significantly higher than those from previous studies.

II. Background

This study focuses on estimating the energy consumption of the storage devices.

Power of storage devices The power of an electronic device is the rate at which electric energy is consumed by the device. The power of different storage devices can be drastically different. An HDD has one or more rotating platters and heads, and a constant power supply is needed to keep the platters spinning and to move the heads to the desired location for each I/O request. An SSD does not need any moving parts. Thus, in general, an HDD's average power is higher than an SSD.

Storage devices have several power management states. After being powered on and starting accepting incoming commands, they are in the "active/idle state". When not needed, they can also be put into a power-saving "standby state," in which the HDD's electric motors are shut off and can save a considerable amount of energy. The drawback is that a device in standby state cannot handle any I/O command before they are bringing back to the active/idle state, and this state change needs several seconds and consumes extra energy to rotating up these platters. A SSD supports similar power management states, but since a SSD has no moving parts to rest, a well-designed SSD should always be in the power-saving mode when not processing requests.

I/O workloads In practice, I/O workloads always require the storage system to provide certain capabilities to run, such as speed, latency, or space. Among these factors, often one is dominant. We classify a workload as a capability workload when its demand for I/O bandwidth is more difficult to meet than its demand for storage space. A web server's workload, for example, is a capability workload when the number of devices needed to meet its bandwidth requirement is larger than the number of devices needed to meet its storage space requirement. The other kind of workload is the capacity workload, whose requirement for storage space dominates over

1526-7539/14 $31.00 © 2014 IEEE
DOI 10.1109/MASCOTS.2014.20

other requirements. Workloads on archival systems are primarily capacity workloads.

The amount of work in a workload can be specified in many ways. Some workloads cover a fixed amount of I/O work, such as importing or exporting a fixed amount of data from the system, file system and file integrity checking, and checkpointing. Some workloads need to run for a certain period, such as a web/file/database server (often generated by a reactive program, which is a process that continuously interacts with the environment, following a pace set by the environment), audio/video recording or streaming, and I/O workloads generated by user-interacting programs (such as text or image processing). Workloads in this category always need to run for the specific amount of time regardless of how fast the storage system is.

III. Models and Metrics

The number of devices needed, N, can be calculated from the workload's capability or capacity requirement. For simplicity, we assume that all devices in the storage system are of the same model, and that no extra devices are needed for providing reliability; their effects on N can be factored in later if needed. Let us consider a workload with peak bandwidth requirement S_{pb}, average bandwidth requirement S_{ab}, and capacity requirement S_c. Let D_b be the bandwidth one storage device can provide and D_c be the storage space one device can provide. Then the number of storage devices needed by the said workload can be calculated as:

$$N = \left\lceil \max\left(\frac{S_{pb}}{D_b}, \frac{S_c}{D_c}\right) \right\rceil \tag{2}$$

Power management schemes also affect the energy consumption. For general-purpose storage systems, the most widely deployed scheme is to put a device into standby after a certain period of inactivity. This policy is conceptually simple, yet in practice it is not immediately clear how to choose the length of inactivity time. For an HDD, state switching is not free, because it involves spinning-up and down the platters, which can consumes a considerable amount of energy. The minimum length of inactivity to justify a state switching is normally called the *break-even* time. There are studies on how to choose the break-even time using statistic models [11]. In practice, this policy is rarely used on capability workloads because often there are not long enough gaps among the incoming requests for the state switch to kick in [5]. For capacity workloads, the common practice is to keep all devices in active state, trying to finish the work as quickly as possible, then put all devices into a standby (or offline) state. These policies are not the only choices, and there are more complex algorithms that are fine-tuned for specific workloads to achieve more aggressive power saving, such as [3]. Ideally, using the optimal power management scheme, only the minimum number of devices that are required to provide the required bandwidth should be kept active and all other devices should be kept in standby or offline state. This optimal scheme might be hard to achieve in practice. Let N_{min} be the theoretical minimum number of devices needed and $\alpha \times N_{min}$ be the actual number of devices that are active. $\alpha \in [1, N/N_{min}]$ is called the power management efficiency factor. An optimal power management scheme should have an α of 1. In the following analyses, we assume that α is 1 for simplicity; a more realistic value can be derived from the users' experience about the domain workloads.

Table I lists the letters used in this paper. D_c is readily available for a given device. The power consumptions in different modes usually need to be measured on the real device because the numbers from the device's specification may not be accurate. D_b and P_b also depend on the characteristics of the workload, such as the ratio of read to write and I/O block size, and have to be measured for each workload/device combination.

TABLE I. List of letters.

Letter	Definition
S_{pb}	A workload's peak bandwidth requirement
S_{ab}	A workload's average bandwidth requirement
S_c	A workload's capacity requirement
D_c	A device's capacity
D_b	A device's bandwidth when running the specific workload
P_i	A device's idle power consumption
P_s	A device's standby power consumption
P_b	A device's busy power consumption when running the specific workload
E	The energy consumption of a workload
T	The run time of the workload
N	The number of devices needed by the workload
α	Power management efficiency factor

A. Energy Consumption Model of Capability Workloads

For capability workloads, we have $S_{pb}/D_b > S_c/D_c$, therefore $N = \lceil S_{pb}/D_b \rceil$. To achieve the required aggregated bandwidth using multiple devices, the common practice is to distribute incoming I/O requests evenly to many devices. The average number of active devices is $\alpha \times S_{ab}/D_b$. We calculate the total energy consumption for running this workload by summing up the energy used by idle devices and the energy used by active devices.

$$\begin{aligned} E &= T \times P_i \times \left(N - \alpha \times \frac{S_{ab}}{D_b}\right) + T \times P_b \times \alpha \times \frac{S_{ab}}{D_b} \\ &= T\left(\frac{P_i}{D_b}(S_{pb} - S_{ab}) + \frac{P_b}{D_b} S_{ab}\right) \end{aligned} \tag{3}$$

Equation (3) expresses the energy consumption in terms of P_i/D_b and P_b/D_b, which are properties of a device. This equation makes it possible to compare the device energy use for running a specific workload by comparing these properties. These properties can also be gotten from the device's specification. In Section V, we measure these properties of several devices and compare them with the values from their specifications.

Random I/O workloads often have I/O requests that are spread across all devices. To handle them, we may have to keep all devices up if the data is evenly spread and the gaps between incoming requests are not long enough to justify a state switching [5]. With all devices active, the energy consumption would be:

$$\begin{aligned} E &= T \times N \times P_b \\ &= T \times S_{pb} \times \frac{P_b}{D_b} \end{aligned} \tag{4}$$

B. Energy Consumption Model of Capacity Workloads

For capacity workloads, we have $S_{pb}/D_b < S_c/D_c$, therefore $N = \lceil S_c/D_c \rceil$. The time needed ($T$) to finish this workload is $T = S_c/S_{ab}$. In practice, S_{ab} may be limited by many factors, such as network speed or source storage media's speed. Very often we have $S_{ab} \ll S_c$, and the whole workload may need from several hours to several days to finish. For capacity workloads, the energy consumption can be calculated as:

$$\begin{aligned} E &= T \times P_i \times \left(N - \alpha \times \frac{S_{ab}}{D_b} \right) + T \times P_b \times \alpha \times \frac{S_{ab}}{D_b} \\ &= \frac{P_i}{D_c} \times \frac{S_c^2}{S_{ab}} + \left(\frac{P_b - P_i}{D_b} \right) \times S_c \qquad (5) \end{aligned}$$

Equation (5) calculates the energy consumption in terms of the following device's properties: P_i/D_c and $(P_b - P_i)/D_b$.

As discussed above, if all devices need to be kept active for running random I/O workloads, we would have:

$$E = \frac{S_c^2}{S_{ab}} \times \frac{P_b}{D_c} \qquad (6)$$

IV. Measurement Setup

We measure a few devices to demonstrate the models. The test machine is a Dell XPS 710 (one dual-core Intel Core 2 processor at 1.86 GHz, 4 GB RAM). The storage devices are listed in Table II. To measure the DC energy consumption, we use the WU100 Version 2 Digital DC Ammeter, Amp Hour & Watt Hour Meter from RC Electronics. It provides 0.001 Ah resolution. We attach one meter to each of the +12 V and +5 V power supply and sum their readings to get the overall energy consumption of the load.

We use the Linux-based operating system Ubuntu LTS 12.04 running in x86 mode. The kernel version is 3.2.0-pae. We do not use a file system, because we are measuring the best that the hardware devices can offer, or in other words, the lower limit of energy consumption when you use a perfect file system that has zero overhead. Without the file system, all I/O requests are sent to the raw device directly and are not affected by any OS cache. We use the `dd` command for generating the sequential workloads and `fio` version 2.1.6.1 for generating the other workloads. All random workloads in this paper are generated according to the uniform distribution. Each experiment is run three times, and the average and variance are calculated.

We evaluate the following essential workloads: sequential read, sequential write, random read, and random mixed read/write (1:1). They are evaluated as both capacity and capability workloads, using different assumptions ($S_{pb}/D_b > S_c/D_c$ or $S_{pb}/D_b < S_c/D_c$) in different categories. All workloads use 1 MB per I/O request. We run each test workload for 30 minutes and record the amount of energy consumed by the specific disk drive that handled the workload.

V. Results

A. Energy Use of Devices

We pick the latest (as of March 2014) high-density, energy-efficient devices from both the 3.5-inch HDD and SSD camps: Seagate Desktop 4 TB HDD ST4000DM000 and Samsung 840 EVO 1 TB SSD. They are both high in storage capacity, density (in terms of TB/$), and power efficiency, but are relatively slow in speed. First, we measure the essential power of the storage devices, which includes their power consumption in various states. The results are listed in Table II. We also include StorageReview's measurement of the latest 2.5-inch high-density low-power HDD, Samsung Spinpoint M9T 2 TB [21], for reference.

The energy needed for doing a power-cycle for HD1 is measured to be 68.58 J. We spin-up, spin-down the device seamlessly for 20 times and measure the total energy consumption. This number is significantly higher than the measurements from previous studies, which were about 6 to 7 J [11]. Our measurement of the power cycle time is on par with previous studies, at about 6 seconds. To further look into this issue, we measure two other HDDs we possess: Seagate Barracuda LP 2 TB (72.94 J) and Seagate Constellation.2 2.5-inch ST9500620NS 500 GB (21.96 J). Their results are all much higher than published results. We will further research this issue, but currently our measurements are more plausible because the power consumption during the spinning-up phase should be higher than the power of idle state because increasing momentum should need more power than maintaining momentum. If the idle power of a HDD is 5 W, the energy consumed in the 5 seconds of spinning-up should be higher than $5 \times 5 = 25$ J. Our findings call into question the previous studies that believe aggressive spin-up and -down is beneficial.

B. Energy Use of Workloads

1) Sequential write workload: The measurements of the sequential write workloads are shown in Table III, which lists all the important properties for calculating the sequential write as either a capability workload or a capacity workload. The columns that have *sp* in their names list the values from the manufacturer's specification. We observe that, for HD1, accessing the head zone and tail zone show significantly differences in speed and energy consumption: accessing the head zone is 49% faster than accessing the tail zone, but also uses 8.8% more Watts. This phenomenon might be caused by the fact that the HDD's platters are always rotating at a fixed speed and the distance covered by a head on the outer platter areas in a second is much longer than on the inner areas.

As shown in Equation (5), the energy use of capability workloads consists of two parts. The first part is proportional to P_i/D_b, the idle power of the device, and $S_{pb} - S_{ab}$, the difference between the peak bandwidth requirement and the average bandwidth requirement. This means that if the bandwidth requirement of the workload has a high variance, the device's idle power would play a very important role in the workload's energy consumption; if the workload's bandwidth variance is low, the idle power would not be very important. The second part is proportional to P_b/D_b. Overall, a more energy efficient device should have lower power usage and higher bandwidth. From Table III, we can see that the P_i/D_b of SSD1 is only about 2.4% of HD1, and the P_b/D_b of SSD1 is only about 34.3% of HD1. This means that SSD1 is more energy efficient than HD1 at running workloads that have a higher peak bandwidth to average bandwidth ratio, like many server loads [3].

Next, we consider the sequential write as a capacity workload, which, according to Equation (5), depends on P_i/D_c

TABLE II. ENERGY USE OF DEVICES.

Drive	Capacity	Cache	$/TB	Standby power (W)	Active/idle power (W)	Power cycle time (s)	Power cycle energy (J)	Misc.
HD1: Seagate Desktop HDD ST4000DM000	4 TB	64 MB	38.75	0.55 (±2.6%)	5.16 (±3.2%)	6.01 (±0.1%)	68.58	[17]
HD2: Samsung 2.5” Spinpoint M9T	2 TB	32 MB	58.5	N/A	0.87	N/A	N/A	5400 RPM [18, 21]
SSD1: Samsung 840 EVO	1 TB	1 GB	505	0.19 (±0%)	0.19 (±4%)	0.002 (±0%)	0	[16]

TABLE III. MEASUREMENT OF DEVICES’ PROPERTIES WHEN RUNNING THE SEQUENTIAL WRITE WORKLOAD. COLUMNS WITH *sp* SHOW THE VALUE FROM THE DEVICE’S SPECIFICATION. D_b IS IN MB/S. P_b AND P_i ARE IN WATTS. P_i/D_b, P_b/D_b, AND $(P_b-P_i)/D_b$ ARE IN 10^{-3}W · S/MB. P_i/D_c IS IN W/TB.

Drive	D_b	$(D_b)_{sp}$	P_b	$(P_b)_{sp}$	P_i	$(P_i)_{sp}$	$\frac{P_i}{D_b}$	$(\frac{P_i}{D_b})_{sp}$	$\frac{P_i}{D_c}$	$(\frac{P_i}{D_c})_{sp}$	$\frac{P_b}{D_b}$	$(\frac{P_b}{D_b})_{sp}$	$\frac{(P_b-P_i)}{D_b}$	$(\frac{(P_b-P_i)}{D_b})_{sp}$
HD1 (head)	155 (±0%)	146	5.45 (±0%)	7.5	5.16 (±3%)	5	33.1	34.2	1.29	1.25	35	51	1.9	17.1
HD1 (tail)	104 (±0%)		5.01 (±0%)				49.6				48		−1.4	
HD2 [21]	124	169	2.86 (±0%)	2.3	0.87	0.7	7	4.1	0.87	0.7	23	14	16	9.4
SSD1	227 (±0%)	520	2.80 (±1%)	0.24	0.19 (±4%)	0.14	0.8	0.3	0.19	0.14	12	1	11.5	0.2

TABLE IV. MEASUREMENT OF DEVICES’ PROPERTIES WHEN RUNNING THE SEQUENTIAL READ WORKLOAD. COLUMNS’ DEFINITION AND UNITS ARE SAME AS ABOVE.

Drive	D_b	$(D_b)_{sp}$	P_b	$(P_b)_{sp}$	P_i	$(P_i)_{sp}$	$\frac{P_i}{D_b}$	$(\frac{P_i}{D_b})_{sp}$	$\frac{P_i}{D_c}$	$(\frac{P_i}{D_c})_{sp}$	$\frac{P_b}{D_b}$	$(\frac{P_b}{D_b})_{sp}$	$\frac{(P_b-P_i)}{D_b}$	$(\frac{(P_b-P_i)}{D_b})_{sp}$
HD1 (head)	152 (±4%)	146	5.79 (±1%)	7.5	5.16 (±3%)	5	33.8	34.2	1.29	1.25	38	51	4.1	17.1
HD1 (tail)	104 (±0%)		5.38 (±2%)				49.6				52		2.1	
SSD1	263 (±0%)	540	1.93 (±0%)	0.24	0.19 (±4%)	0.14	0.7	0.3	0.19	0.14	7	0.4	6.6	0.2

TABLE V. MEASUREMENT OF DEVICES’ PROPERTIES WHEN RUNNING THE RANDOM READ WORKLOAD. COLUMNS’ DEFINITION AND UNITS ARE SAME AS ABOVE. P_b/D_c IS IN W/TB. SSD1’S SPECIFICATION DOES NOT CONTAIN AN OFFICIAL BANDWIDTH FOR THIS OR SIMILAR WORKLOAD.

Drive	D_b	$(D_b)_{sp}$	P_b	$(P_b)_{sp}$	P_i	$(P_i)_{sp}$	$\frac{P_i}{D_b}$	$(\frac{P_i}{D_b})_{sp}$	$\frac{P_i}{D_c}$	$(\frac{P_i}{D_c})_{sp}$	$\frac{P_b}{D_b}$	$(\frac{P_b}{D_b})_{sp}$	$\frac{P_b}{D_c}$	$(\frac{P_b}{D_c})_{sp}$
HD1	39.28 (±0%)	146	5.26 (±0%)	7.5	5.16 (±3%)	5	131.3	34.2	1.29	1.25	134	51	1.31	1.875
SSD1	268 (±0%)	N/A	1.96 (±2%)	N/A	0.19 (±4%)	0.14	0.7	N/A	0.19	0.14	7	N/A	1.96	N/A

TABLE VI. MEASUREMENT OF DEVICES’ PROPERTIES WHEN RUNNING THE RANDOM READ/WRITE WORKLOAD. COLUMNS’ DEFINITION AND UNITS ARE SAME AS ABOVE. SSD1’S SPECIFICATION DOES NOT CONTAIN AN OFFICIAL BANDWIDTH FOR THIS OR SIMILAR WORKLOAD.

Drive	D_b	$(D_b)_{sp}$	P_b	$(P_b)_{sp}$	P_i	$(P_i)_{sp}$	$\frac{P_i}{D_b}$	$(\frac{P_i}{D_b})_{sp}$	$\frac{P_i}{D_c}$	$(\frac{P_i}{D_c})_{sp}$	$\frac{P_b}{D_b}$	$(\frac{P_b}{D_b})_{sp}$	$\frac{P_b}{D_c}$	$(\frac{P_b}{D_c})_{sp}$
HD1	44.28 (±0%)	146	5.16 (±0%)	7.5	5.16 (±3%)	5	116.5	$342e^{-4}$	1.29	1.25	116	51	1.29	1.88
SSD1	225 (±3%)	N/A	2.33 (±0%)	N/A	0.19 (±4%)	0.14	0.8	N/A	0.19	0.14	10.4	N/A	2.33	N/A

and $(P_b - P_i)/D_b$. We know $T = S_c/S_{ab}$, Equation (5) can be transformed to:

$$E = S_c\left(\frac{P_i}{D_c} \times T + \frac{P_b - P_i}{D_b}\right)$$

Using the measured values of P_i/D_c and $(P_b - P_i)/D_b$ from Table III, we can calculate the condition for SSD1 to use less energy than HD1 when running a sequential write workload:

$$\begin{aligned} & E_{HD1} > E_{SSD1} \\ \Rightarrow \quad & 1.29 \times 10^{-6}T + 1.9 \times 10^{-3} > 0.19 \times 10^{-6}T + 11.5 \times 10^{-3} \\ \Rightarrow \quad & T > 8727.27\,\mathrm{s} \end{aligned}$$

It shows that SSD1 is more energy efficient than HD1 when the size of the data is more than 8727 times of the aggregated write bandwidth. The unit of T is a second. 8727 seconds is about 2.4 hours. This means that if a sequential write workload requires more than 2.4 hours to run, using SSD1 would be more energy efficient. On the HDD side, since this is a sequential workload, if we can optimize the power management scheme so that all unneeded devices are put into standby mode (instead of idle), we can change Equation (5) to:

$$\begin{aligned} E &= T \times P_s \times \left(N - \alpha \times \frac{S_{ab}}{D_b}\right) + T \times P_b \times \alpha \times \frac{S_{ab}}{D_b} \\ &= \frac{P_s}{D_c} \times \frac{S_c^2}{S_{ab}} + \left(\frac{P_b - P_s}{D_b}\right) \times S_c \end{aligned} \tag{7}$$

Using Equation (7) to compare HD1 and SSD1:

$$\begin{aligned} & E_{HD1} > E_{SSD1} \\ \Rightarrow \quad & 0.14 \times 10^{-6}T + 32.2 \times 10^{-3} > 0.19 \times 10^{-6}T + 11.5 \times 10^{-3} \\ \Rightarrow \quad & T < 414000\,\mathrm{s} \end{aligned}$$

This means that using HD1 is more energy efficient than using SSD1 for sequential write workloads that need more than $414000/60/60 = 115$ hours to run, if all unused HDDs can be put into standby mode.

Here we ignore a fact that you cannot power off a fraction of a HDD. That is to say, if you have a workload that writes 12 TB data at 200 MB/s, you cannot turn 1/3 HDD off to save energy. The discussion above applies when a large number of devices are involved and their statistical aggregated behavior conforms to the equations above.

Considering the cost, as of March 2014, HD1’s retail price is $38.75 per TB and SSD1’s retail price is $505 per TB, and California's electricity price is about \$0.4 per kWh for commercial usage. Using Equation (5), we can get:

$$\frac{38.75}{10^6} + \frac{0.4(1.29 \times 10^{-6}T + 1.9 \times 10^{-3})}{1000 \times 3600} > \frac{505}{10^6} + \frac{0.4(0.19 \times 10^{-6}T + 11.5 \times 10^{-3})}{1000 \times 3600}$$

Solving it, we get $T > 3.8 \times 10^9$ second, which means that the workload's total data volume must be prohibitively larger than its writing bandwidth for SSD1 to cost less than HD1. In other words, the energy saved by using SSD1 will never justify the high purchase price of the device. This proves that using SSD will never be cost effective to run the sequential write workload as described in this section.

We do not have the latest 2.5" low power high density HDD in possession, but this should not stop us from doing a thought experiment using data of Samsung Spinpoint M9T 2 TB (HD2) [18] from [21]. For HD2 to be more energy efficient than HD1, we can solve the inequality and get $T > 33571.4$, which means sequential write workloads that need more than 9.3 hours to run. When the costs of devices are included, we need $T > 4.2 \times 10^8$ for HD2 to be more cost effective than HD1, which is also prohibitively large. Thus, we can see that the cost advantage of high-density low-cost HDDs, like HD1, is hard to beat when running such a capacity workload.

Comparing our measurements with the device's specification, we can see that HD1's performance figures from its specification are on par with our measurements. On the contrary, SSD1's measured performance differs notably from its specification, especially for the bandwidth D_b and busy power P_b, which is 56.3% lower and 10.7 times higher respectively.

2) Sequential read workload: The results of the sequential read workload are shown in Table IV. Using the same analytical method, we can see that SSDs are more energy efficient for the capability sequential read workload. For the capacity sequential read workload, it is more energy efficient to use SSDs when $T > 2273$ (if unused HDDs are left in idle) or $T < 548000$ (if unused HDDs are put in standby state), using the same inequalities as discussed above. Likewise, if the device purchase costs were included, SSDs would not be cost efficient.

3) Random read workload: The results of the random read workload are shown in Table V. When we consider the random read workload as a capability workload, Equation (4) shows that devices with lower P_b/D_b should be favored. But when considering the workload as a capacity workload, using Equation (4), we can see that the energy consumption is proportional to P_b/D_c, and our measurements show that HD1 has a lower P_b/D_c than SSD1, which means that HD1 is actually more energy efficient. For example, when we need to server 4 TB of data for random read, as a capacity workload, the bandwidth requirement is not critical so we can use either one HD1 or four SSD1. Using one HD1 needs 5.26 W, and using four SSD1 needs $1.96 \times 4 = 7.84$ W. Clearly, here the HD1 is not only a cheaper choice, but also a greener choice.

4) Random read/write (1:1) workload: The results of the random read/write (1:1) workload are shown in Table VI. Unsurprisingly, the analytical results are similar to the random read workload: SSD1 is more energy efficient when the random read/write workload is a capability workload, and HD1 is more energy efficient when the workload is a capacity workload.

Very often, a device is not running at full speed, such as when running capacity workloads. To see whether running a device at half speed can lead to energy saving, we measure a device's energy consumption when its bandwidth is capped at a half. For HD1, the half-speed power is 98% of full-speed power. For SSD1, the half-speed power is 86% of full-speed power. None of them scales linearly when running at half speed. Thus, it is not energy efficient to keep devices running at half speed. Instead, an optimal power management scheme should keep as few devices up as possible, running them at full speed, and put all unused devices to idle.

VI. Conclusion

The models in this paper bridge the gap between the energy use of workload and the energy use of storage devices. Using several key performance properties of the workload, we can use these models to compare the energy consumption of storage devices and to pick the greenest or lowest cost option. We measure the two latest high-density storage devices to illustrate our idea under various common workloads. We cannot say whether SSDs are more energy efficient than HDDs until enough data about the specific workload and devices are collected. As an educated guess, SSDs are probably more energy efficient for running capability workloads, but not for capacity workloads. However, SSDs are rarely cost efficient. The situation would be even worse if we were to consider the fact that using more SSDs requires more ports on the server, which in turn pushes up the number of servers required.

We raise the question about the energy needed for a HDD to do one power-cycle because our measurements are notably higher than previous studies. We also discover that the energy use of a storage device is not proportional to its utilization. Therefore, to maximize energy efficiency, active devices should be kept at 100% busy, and unused devices should be put into idle. This applies to both HDDs and SSDs.

The models are limited in several ways. First, we assume that the file system has zero overhead and the data placement algorithm is optimal. Therefore, our models calculate the theoretical lower limit of the device energy use. Second, we assume that the storage system is homogeneous and all its storage devices are of the same model. It is possible to further optimize the system by using a combination of different devices. For these more complex situations, the core idea, process, and conclusions of this paper still apply.

Our measurement results are limited by the number of devices, but the large differences between different types of devices merit attention from device manufacturers, system designers, and end users. We hope to see more analyses that survey a wide range of devices or tackle the needs of specific user scenarios using the methodology from this paper. Our test scripts and raw data are published online at: https://bitbucket.org/ssrc/modeling-workload-energy.

VII. Related Work

Recent models of the energy use of general storage devices include [2, 13, 20, 23]. Other studies look into specific scenarios or settings, like embedded/mobile platforms [10, 13], servers [14, 15], data centers [4], HPC system [7], Big Data processing [6], archival storage system [1, 22], erasure-coding [9], and interaction between file system, devices and workload [8, 19]. From the power management point of view, these studies rarely distinguish capacity workloads from capability workloads and do not consider the number of devices or time needed to run a specific workload.

Acknowledgment

This research was supported in part by the National Science Foundation under awards IIP-1266400, CCF-1219163, CNS-1018928, the Department of Energy under award DE-FC02-10ER26017/DESC0005417, and industrial members of the Center for Research in Storage Systems. We would like to thank the sponsors of the Storage Systems Research Center (SSRC), including Department of Energy/Office of Science, EMC, Hewlett Packard Laboratories, Hitachi, Huawei, Intel Corporation, LSI, NSF, NetApp, Permabit Technology, Samsung Information Systems America, SanDisk, and Seagate Technology for their generous support. We also thank the faculty and students of SSRC, especially Zhichao Hu, Christina Strong, Ethan Miller, Thomas Marlette, and DJ Capelis.

References

[1] I. Adams, E. L. Miller, and M. W. Storer. Examining energy use in heterogeneous archival storage systems. In *Proceedings of the 18th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '10)*, pages 297–306, Aug. 2010.

[2] M. Allalouf, Y. Arbitman, M. Factor, R. I. Kat, K. Meth, and D. Naor. Storage modeling for power estimation. In *Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference*, SYSTOR '09, pages 3:1–3:10, New York, NY, USA, 2009. ACM.

[3] R. Bianchini and R. Rajamony. Power and energy management for server systems. *Computer*, 37(11):68–76, Nov. 2004.

[4] T. Bostoen, S. Mullender, and Y. Berbers. Power-reduction techniques for data-center storage systems. *ACM Computing Surveys*, 45(3):33:1–33:38, July 2013.

[5] E. V. Carrera, E. Pinheiro, and R. Bianchini. Conserving disk energy in network servers. In *Proceedings of the 17th International Conference on Supercomputing (ICS '03)*, pages 86–97, New York, NY, USA, 2003. ACM.

[6] A. M. Caulfield, L. M. Grupp, and S. Swanson. Gordon: Using flash memory to build fast, power-efficient clusters for data-intensive applications. In *Proceedings of the 14th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, Mar. 2009.

[7] M. L. Curry, H. L. Ward, G. Grider, J. Gemmill, J. Harris, and D. Martinez. Power use of disk subsystems in supercomputers. In *Proceedings of the 6th Parallel Data Storage Workshop (PDSW '11)*, Nov. 2011.

[8] D. Daniel and J. Belak. Topic 5: Resiliency and file system I/O (and power!). DOE SC/ASCR and NNSA/ASC Exascale Research Conference, Apr. 2012.

[9] K. Greenan, D. D. E. Long, E. L. Miller, T. Schwarz, and J. Wylie. A spin-up saved is energy earned: Achieving power-efficient, erasure-coded storage. In *Proceedings of the 4th Workshop on Hot Topics in System Dependability (HotDep '08)*, Dec. 2008.

[10] D. P. Helmbold, D. D. E. Long, and B. Sherrod. A dynamic disk spin-down technique for mobile computing. In *Proceedings of the 2nd Annual International Conference on Mobile Computing and Networking 1996 (MOBICOM '96)*, pages 130–142, Rye, New York, Nov. 1996. ACM.

[11] S. Irani, S. Shukla, and R. Gupta. Competitive analysis of dynamic power management strategies for systems with multiple power saving states. In *Proceedings of the 2002 Design, Automation Test in Europe Conference Exhibition (DATE '02)*, pages 117–123, 2002.

[12] J. G. Koomey. Growth in data center electricy use 2005 to 2010. Technical report, Stanford University, Aug. 2011.

[13] J. Li, A. Badam, R. Chandra, S. Swanson, B. Worthington, and Q. Zhang. On the energy overhead of mobile storage systems. In *Proceedings of the 12th USENIX Conference on File and Storage Technologies (FAST)*, pages 105–118, Berkeley, CA, 2014. USENIX.

[14] D. Narayanan, E. Thereska, A. Donnelly, S. Elnikety, and A. Rowstron. Migrating server storage to SSDs: Analysis of tradeoffs. In *Proceedings of the 4th ACM European Conference on Computer Systems (EuroSys '09)*, pages 145–158, New York, NY, USA, 2009. ACM.

[15] E. Pinheiro and R. Bianchini. Energy conservation techniques for disk array-based servers. In *Proceedings of the 18th International Conference on Supercomputing (ICS '04)*, June 2004.

[16] Samsung Electronics. 840 EVO-Series 1TB 2.5-Inch SATA III Single Unit Version Internal Solid State Drive MZ-7TE1T0BW. http://www.samsung.com/us/computer/memory-storage/MZ-7TE1T0BW.

[17] Seagate Technology LLC. Desktop HDD data sheet. http://www.seagate.com/www-content/product-content/barracuda-fam/desktop-hdd/barracuda-7200-14/en-gb/docs/desktop-hdd-data-sheet-ds1770-1-1212gb.pdf.

[18] Seagate Technology LLC. Spinpoint M9T Mobile SATA Drive. http://www.seagate.com/www-content/support-content/samsung/internal-products/spinpoint-m-series/en-us/samsung-m9t-internal-ds.pdf.

[19] P. Sehgal, V. Tarasov, and E. Zadok. Evaluating performance and energy in file system server workloads. In *Proceedings of the 8th USENIX Conference on File and Storage Technologies (FAST)*, Feb. 2010.

[20] A. L. Shimpi. Samsung SSD 840 EVO Review: 120GB, 250GB, 500GB, 750GB & 1TB Models Tested. http://www.anandtech.com/show/7173/samsung-ssd-840-evo-review-120gb-250gb-500gb-750gb-1tb-models-tested/11, July 2013.

[21] StorageReview.com. Samsung Spinpoint M9T hard drive review. http://www.storagereview.com/samsung_spinpoint_m9t_hard_drive_review.

[22] M. W. Storer, K. M. Greenan, E. L. Miller, and K. Voruganti. Pergamum: Replacing tape with energy efficient, reliable, disk-based archival storage. In *Proceedings of the 6th USENIX Conference on File and Storage Technologies (FAST)*, Feb. 2008.

[23] B. Yoo, Y. Won, S. Cho, S. Kang, J. Choi, and S. Yoon. SSD characterization: From energy consumptions perspective. In *Proceedings of the 3rd Workshop on Hot Topics in Storage and File Systems (HotStorage '11)*, 2011.