Energy Proportionality for Storage: Impact and Feasibility

Jorge Guerra[†], Wendy Belluomini[*], Joseph Glider[*], Karan Gupta[*], Himabindu Pucha[*]
jguerra@cs.fiu.edu, {wb1,gliderj,guptaka,hpucha}@us.ibm.com
[†]*Florida International University, IBM Almaden Research Center*[*] *(in alphabetical order)*

Abstract

This paper highlights the growing importance of storage energy consumption in a typical data center, and asserts that storage energy research should drive towards a vision of energy proportionality for achieving significant energy savings. Our analysis of real-world enterprise workloads shows a potential energy reduction of 40-75% using an ideally proportional system. We then present a preliminary analysis of appropriate techniques to achieve proportionality, chosen to match both application requirements and workload characteristics. Based on the techniques we have identified, we believe that energy proportionality is achievable in storage systems at a time scale that will make sense in real world environments.

1 Introduction

Energy consumption of data centers is a significant portion of their operational cost, and has become a major concern both for their owners and the countries in which they are placed. Thus, a large amount of recent work has focused on improving the energy efficiency of data centers in all aspects—distribution of power, data center cooling, and energy consumption of IT components. Among these different aspects, reducing the energy consumption of IT components (servers, storage, networking equipment etc.) plays an important role since improving its energy efficiency also reduces the required capacity of the cooling and power distribution systems.

Of all the IT components in a data center, servers have been the dominant energy consumers. While a single-core microprocessor in 2005 consumed 100 W of energy [1], a disk consumed around 15 W. Thus, server energy consumption attracted significant research attention, and typical server energy usage has decreased considerably over the last few years. Today an idling (power-gated) core consumes as little as 3 W of energy and an active core consumes around 20 W [2]. On the other hand, disk drive power consumption has remained relatively unchanged per drive. As a result, storage is consuming an increasing percentage of energy in the data center. Recent work has shown that in a typical data center today, storage[1] accounts for up to 37-40% of the energy consumption of all IT components [3, 14].

We expect storage energy consumption to continue increasing in the future as data volumes grow and disk performance and capacity scaling slow. A recent study by IDC [13] makes the following observations that back this trend—(1) storage unit acquisition will likely outpace server unit acquisition during the next five years; (2) storage capacity per drive is increasing more slowly, which will force the acquisition of more drives to accommodate growing capacity requirements; (3) data centers are predicted to move towards 2.5” drives that typically consume more energy per GByte than their 3.5” equivalents; and (4) performance improvements per drive have not and will not keep pace with capacity improvements. Thus, improving IOPS per watt continues to be a challenge. Attesting to this growing importance of storage energy consumption, EPA announced EnergyStar specifications for storage components in April 2009.

A rich body of existing work (e.g., [7, 8, 18, 12, 19, 17, 9]) has already investigated energy efficiency of storage systems. However, a significant fraction of this work assumes the existence of hard disks with Dynamic RPM capability (e.g., [8, 18, 12, 19]). However, DRPM drives are not being commercialized in quantity by any major drive vendor due to physical and cost constraints. Nevertheless, the increasing importance of storage energy has spurred innovations in hard disk design such as multiple idle modes and just-in-time seeks, and Solid State Disks (SSDs) are poised to improve IOPS per watt dramatically. In light of this, we argue that the research community should renew interest in improving storage energy consumption at the storage subsystem and the data center level. More importantly, we claim that improving energy efficiency alone is not adequate, and that significant efforts must be focused on achieving *energy proportionality* for storage systems.

Energy proportionality was first proposed for servers by Barroso and Holzle [5]. This work observed that 5000 servers, over a six month period, spent most of their time between 10 and 50 percent utilization. Thus, the authors argued that energy usage should vary as utilization levels vary. Specifically, energy proportionality suggests that as the amount of work done increases, so can the energy consumed to perform it.

This paper investigates whether this concept of energy proportionality can and should be extended to storage. We argue that an energy proportional storage system is useful in two different scenarios:
When performance matters most—storage energy consumption should vary with the performance requirement. This perspective is important for normal operation of a data center.
When energy matters most—performance provided by the storage should be regulated according to energy constraints. This perspective could be important for operation of data centers that are experiencing a transient (e.g. brownout) or chronic energy constraint.

The first perspective (performance matters most) is more achievable in the short-term, but we believe that the second perspective is important in the long-term, especially as the focus shifts to storage energy management. Ultimately, storage energy management should be part of an integrated data center energy management architecture. Focusing on energy proportionality will enable storage energy management to be coordinated with server, network and cooling energy management.

The rest of our paper makes the following contributions: (1) exploration of the benefit from an energy proportional storage sys-

[1] Storage, in the rest of this paper, refers to the collection of independently managed dedicated clusters/subsystems for storage, and not direct attached storage.

tem. To that end, we first study the variation in performance demands on storage systems in real-world enterprise data center environments and estimate the potential energy savings (Section 2), (2) outlining of techniques that can be used to build energy proportional storage systems, and systematic analysis of these techniques (Section 3.2), and (3) illustrating how storage application requirements and workload characteristics map to storage energy proportionality techniques (Section 3.3.3).

2 Energy Proportionality Matters!

This section motivates the importance of an energy proportional storage architecture by studying the variation in utilization seen by units of storage.

2.1 Methodology

We use two trace sets, collected at volume level and extent (fixed-sized partition of a volume) level respectively, for this study. The volume-level trace was collected from the data center of a financial corporation. The data center contained 10 large storage systems, with a total of 10124 volumes mounted on RAID-5 arrays, each with 7 or 8 hard disks. I/O data (metrics include I/O rate, read-to-write ratio, random-to-sequential ratio, and average transfer size) for each volume in the storage system was collected by a centralized independent server every 15 minutes and stored in a database. This data was later compacted into per-day averages on a per-volume basis. We use these per-day samples, over 160 days, in our study. Apart from using I/O rate to understand the variation in the performance demands, we compute utilization level for each volume as:

$$\frac{\text{Instantaneous I/O rate}}{\text{Max I/O rate for that volume}}$$

The extent-level trace was collected by running a public storage benchmark modeling a transactional workload for an extended time period. The setup consisted of an IBM pSeries computer running AIX connected to an IBM storage system with 14 RAID-5 arrays (each array consisting of 7 or 8 disks). Fourteen 1 TByte volumes were allocated on this storage system, and the benchmark accessed different extents in those volumes. A monitoring module recorded I/O rate periodically for each unique extent on all volumes.

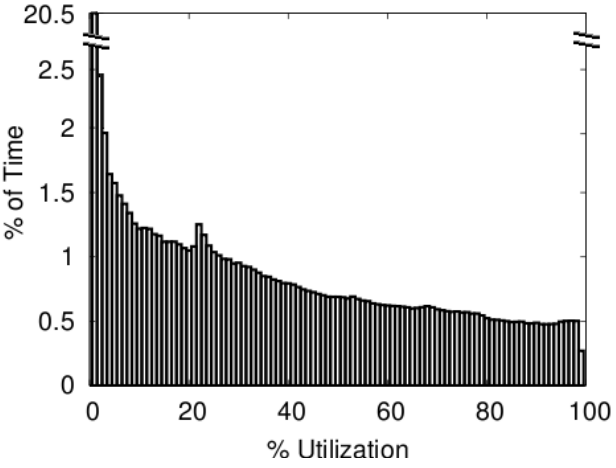

Figure 1: Average fraction of time vs. volume utilization (volume-level trace).

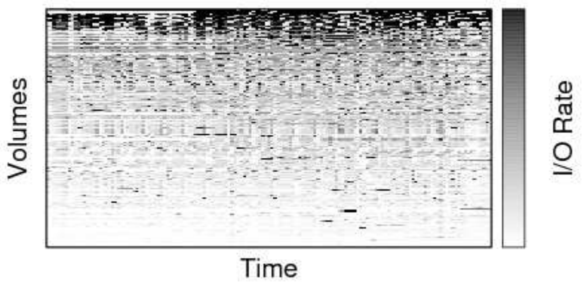

Figure 2: Heatmap depicting I/O rate across volumes for the volume trace. Note that the volumes are sorted using the median I/O rate from bottom to top on the Y-axis. Darker color depicts higher I/O activity.

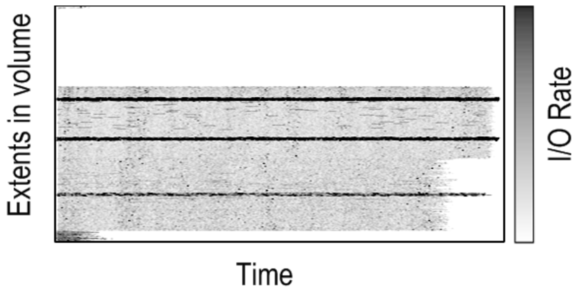

Figure 3: Heatmap depicting I/O rate across extents in a single volume in the extent trace. Darker color depicts higher I/O activity.

2.2 Impact of Energy Proportionality

Figure 1 depicts the average percentage of time a volume spent at each utilization level for the volume level trace. On average, a volume is idle for ∼20% of the time (first two columns). For the remaining 80%, a decreasing amount of time is spent at each utilization level. Clearly, this shows that on average, volume utilization in this storage system is highly variable, with a significant amount of time spent at moderate utilization levels. This highlights the need for storage systems to exhibit energy proportional behavior.

While Figure 1 depicts the average behavior across all volumes in the trace, Figure 2 shows the I/O activity over time for each volume. We see that most volumes have variable usage over time. However, a few volumes do maintain a consistently high or low I/O rate for a majority of the time (top and bottom portions of the figure). Thus, both energy efficiency and proportionality are essential for potential energy savings.

Given the I/O distribution seen in the volume trace, we now estimate the energy savings in an ideally energy proportional system. In such a system, energy usage would follow utilization, which on average is less than 60%. Therefore, the potential for energy savings in an ideally proportional system exceeds 40% when compared to a system always running at 100% utilization (and hence energy) independent of system demands.

To understand whether there are more opportunities at a finer-grain, we also studied an extent-level trace described previously. Figure 3, which is typical of all the volumes in the trace, shows that: (1) some extents experience variable I/O activity over time, while others remain largely idle. (2) a few extents have much higher average I/O activity than other non-idle extents (depicted by the horizontal dark bands). This indicates the potential to achieve proportionality at a granularity finer than entire volumes or arrays. For example, if the most active 4% of extents were resident on enterprise class SSD disks and the remaining 96% on SATA disks, we calculate energy savings of nearly 75% in contrast to a system completely built from enterprise disks, while the acquisition cost of the system would be about the same. Moreover, extents can be easily moved (compared to volumes) to and back from different disk tiers when their utilization changes.

3 Storage Energy Proportionality

Given the significant savings from an ideally energy proportional storage architecture, we now explore various techniques that can be leveraged to achieve proportionality. We begin by understanding the techniques used to achieve server energy proportionality in order to identify analogous approaches for storage if possible.

3.1 Servers vs. Storage Techniques

A broad class of techniques for achieving server energy proportionality are based on Dynamic Voltage and Frequency Scaling (DVFS) (e.g., [6]). DVFS provides lower energy active modes (modes where servers can continue to perform work at lower performance), and allows relatively fast switching between modes. On the other hand, there is currently no powerful analogue in storage that directly enables fine-grain scaling of performance as a tradeoff for energy.

Another class of server techniques focuses on achieving proportionality by migrating workloads such that servers either operate at their peak, or can be suspended [16]. This is in contrast to the above techniques that scale performance, and are considered more powerful since lower performance states are less energy efficient. On the other hand, the time-scale to adapt for these techniques is higher since it involves process migration. Although this class of techniques is directly applicable to storage, the overhead incurred in storage will likely be higher given its statefulness and amount of data moved. Moreover, data availability requirements often imply that data should always be reachable within small number of milliseconds or seconds.

3.2 Storage Techniques

We explore several energy saving techniques, a majority from existing literature, and a few novel techniques inspired by server-side approaches and energy saving hard disk modes. These techniques, applied alone or in concert with each other, can contribute to an energy proportional storage architecture.

The first set of techniques use intelligent data placement and/or continuous data migration, resulting in large time-scale to adapt. Examples include:

Consolidation: Aggregation of data into fewer storage devices whenever performance requirements permit.

Tiering/Migration: Placement/movement of data into storage devices that best fit its performance requirements (e.g. [17]).

Write off-loading: Diversion of newly written data to enable spinning down disks for longer periods, coupled with opportunistic movement of data to storage devices when they become active [9]. The same technique can be used to efficiently service peaks in I/O activity [10].

The next class of techniques benefit from the availability of hardware-based active and inactive low energy modes in disks. As such, the overhead of these techniques is low since it does not involve data movement. Examples include:

Adaptive seek speeds: Allow trading off performance for power reduction by slowing the seek and waiting an additional rotational delay before servicing the I/O [15].

Workload shaping: Batching I/O requests to allow hard disks to enter low power modes for extended periods, or to allow workload mix optimizations [4].

Opportunistic spindown: Spinning down hard disks when idle for a given period.

Spindown/MAID: Maintaining disks with unused data spundown most of the time, either by concentrating "important" data [7, 12] or tuning caching and prefetching algorithms to save power [11, 18], in all cases trying to increase idle periods of disks with low load.

Finally, another class of techniques reduces the amount of space required to store data directly resulting in energy savings. For example, **dedup/compression** involves storing smaller amounts of data using very efficient representations of the data.

3.3 Framework for Technique Selection

Given the above techniques, an important question to answer is which techniques are suited for a given storage application. This section outlines a framework to make such recommendations.

Different techniques present different tradeoffs with respect to their potential to alter application performance, and incurred overhead versus resulting benefit. For example, a technique that leverages opportunistic spindown can potentially cause the peak response time of an application to be as high as the spin up delay (∼10 seconds). Similarly, techniques that rely on data migration incur significant overhead, and hence rely on workloads that amortize the overhead to result in energy savings. Thus, our framework selects appropriate techniques using two inputs: (1) application performance requirements, and (2) application workload characteristics. In the rest of this section, we categorize these inputs based on their interaction with the techniques, and then match the different categories with the appropriate techniques to achieve energy efficiency and proportionality.

3.3.1 Application Performance Requirements

Since the above techniques may impact application performance, their usage must be aligned with the requirements of the storage applications. To understand this better, we propose three different categories of storage applications–although the range of applications probably lies more along a continuum than in discrete categories–taking into account two factors: sensitivity to average response time, and sensitivity to peak response time (e.g. from spin up delay).

1. High sensitivity to peak response time, high sensitivity to average response time: These are often critical business applications typically using SAN storage which have short default timeouts. Transactional databases are an example.

2. Low sensitivity to peak response time, high sensitivity to average response time: These are often but not always important business or consumer applications which require good storage performance but can tolerate occasional delays. Examples include web storage, multimedia streaming and general user file storage.

3. Low sensitivity to peak response time, low sensitivity to average response time: These are often archival/backup applications where multi-second delays are tolerated and response time is not

Table 1: Attributes of techniques for storage energy proportionality.

Technique	App Category	Time-scale	Granularity	Potential to alter performance
Consolidation	1,2,3	hours	coarse	Can lengthen response times
Tiering/migration	1,2,3	minutes-hours	coarse	Can lengthen response times
Write off-loading	2,3	milliseconds	coarse	Adds background process that can impact application
Adaptive seek speeds	1,2,3	milliseconds	fine	Can lengthen response times
Workload shaping	2,3	seconds	fine	Can lengthen response times
Opportunistic spindown	2,3	seconds	fine	Delays due to spinup
Spindown/MAID	3	10's of seconds	medium	Delays due to spinup
Dedup/compression	2,3	n/a	n/a	Delays in accessing data due to assembling from repository or decompression

so important. Examples include medical or generic archival/backup and eDiscovery.

Because performance requirements vary between these categories, different techniques will be appropriate for applications in different categories. For example, it may never be permissible to spindown a hard disk supporting a high peak response time intolerant application whereas opportunistic spindown may be acceptable for an application which is tolerant of high peak response time but generally requires average response time in the tens of milliseconds.

Table 1 presents a summary of which storage techniques can be used by different application categories, as well as the tradeoffs presented by these techniques using the following metrics:

Time-scale to adapt: Some techniques affect the environment in milliseconds (e.g. slowing seek speeds) while others take seconds (e.g. spinning down disks), minutes or hours (e.g. migrating data).

Granularity of control: Some techniques are fine grain (e.g. adjustment of I/O dispatch rate) while others can be coarse grain (e.g. migrating a volume).

Potential to alter performance: Some techniques can introduce small delays to access times (e.g. spinup) while others will not (e.g. consolidating).

3.3.2 Workload Characteristics

The extent of energy savings and the incurred overhead of different techniques is significantly impacted by the workload characteristics of the applications using the storage system.

An important characteristic for data movement-based techniques is the steadiness of the workload; a steady workload exhibits a constant utilization over time. This enables these techniques to place data optimally, and accrue benefits over time. Even if the workload is not perfectly steady, these techniques benefit if the workload shows periodicity. On the other hand, these techniques cannot quickly respond to variation in the utilization of the workload, and can only adapt to average utilization, thereby reducing the benefit. For convenience, we refer to these features as *workload stability*. In fact, given the higher overhead of data movement based techniques, workload stability decides if such techniques are applicable.

Another workload characteristic that affects the amount of benefit from the techniques is their extent of utilization. Workloads with low utilization most of the time can benefit from spindown, whereas workloads with high utilization majority of the time have fewer spindown opportunities.

As an example, we categorize the workloads in the volume-level trace. We first break up the volumes based on their stability and then their average utilization. Overall, we see that almost 30% of the volumes exhibit stability. We then divide these stable volumes based on their utilization. We observe that 10% of the volumes are highly utilized, and have $\geq$ 80% utilization more than 90% of the time. Data in this group can benefit from suitable tiering and consolidation since the required time-scale for change is rather high. Such data can also be colocated with data experiencing low utilization.

Table 2: Volume categorization for the financial data center workload. Key: *H*: high load, *L*: low load, *P*: peaks in load, *V*, V_X: variable load (V_1=lowest, V_4=highest I/O rate).

Category	H	L	P	V	V_1	V_2	V_3	V_4
% Vol.	10	5	13	72	51	6	4	11

Another 5% of the volumes have mostly idle data units, with $\leq$ 20% utilization more than 90% of the time. Such data units can be consolidated and migrated onto slower devices, which may then be spundown if the application category allows for the spinup delay. Write-offloading and workload-shaping techniques are also suitable.

The next 13% of volumes consist of data units that are mostly under utilized but have bursty peaks, nominally with $\leq$ 50% utilization more than 70% of the time and $\geq$ 80% utilized more than 10% of the time. In this case, peaks can be serviced by fine-grain migration which has relatively low overhead. Hard disks which allow slow seeks would also provide a quick-to-change method of aligning I/O performance characteristics and energy usage as I/O load changes. In addition, write-offloading techniques allow for handling of small peaks for write traffic while also allowing spindown of hard disks holding most of the data [10].

The remaining 72% of volumes show a fairly variable workload. We further divide the volumes in this category into four more groups, characterized by their maximum I/O rate in comparison with the maximum I/O rate observed in the system (the last four columns of the Table 2) such that V_1 holds volumes with lowest I/O rate and V_4 holds volumes of highest I/O rate. We see that most of the volumes have low I/O rate (V_1). These need a smaller range of energy-utilization points. Fine grain migrations across different RAID tiers (RAID-5 vs. RAID-10) could potentially be useful. Volumes in V_2-V_4 have the most variability in their I/O rate and therefore need techniques with small time-scale to adapt; slow-seek techniques and fine-grain migrations across device types (SSD/SAS/SATA) would be the most suitable.

3.3.3 Alignment of Techniques

Table 3 expresses the application response requirements in two dimensions: 1) as being sensitive (or not) to lengthening of peak response time, and 2) as being sensitive (or not) to drive spinup delays. It also factors in workload characteristics; we specifically focus on stability since it affects the applicability of a technique, as opposed to the other characteristics that only impact the amount of benefit. For example, stability indicates if an adaptive algorithm will be inef-

Table 3: Framework for mapping storage application performance requirements and workload characteristics to energy saving techniques. Techniques: C: *Consolidation*, T: *Tiering/Migration*, S: *Opportunistic Spin-down/MAID*, W: *Write Offloading*, A: *Adaptive Seek Speeds*, H: *Workload Shaping*, D: *Dedup/Compression*.

Sensitivity to	Sensitivity to	Stability	Techniques						
Avg. Resp. Time	Peak Resp. Time	of Workload	C	T	S	W	A	H	D
Yes	Yes	No							
		Yes	✓	✓					
	No	No			✓	✓			
		Yes	✓	✓	✓	✓			
No	No	No			✓	✓	✓	✓	✓
		Yes	✓	✓	✓	✓	✓	✓	✓

Key ✓ : Applicable.

fective due to unexpected shifts; the more unpredictable the future, the less likely that a high cost/overhead technique will be useful. Given those three parameters, we suggest in the table which techniques are most likely to be effective in which mix of application requirement and workload.

Finally, we also considered whether individual techniques would clash with each other or be complementary. On the whole, we feel that the techniques need not clash, meaning that a number of them may be deployed effectively in the same system. However, techniques that generally are complementary might clash on individual resources or at points in time. A simple example of this would be opportunistically spinning down of some drives while the system was planning to migrate data off the drives to turn them off. Another example might be lengthening seek times on a drive where workshaping algorithms are getting ready to execute a batch of I/O requests. The lesson from this is that as these techniques are incorporated into a system, careful architecture is required to ensure appropriate decisions about usage of the techniques.

4 Conclusion

Given the increasing energy consumption by storage systems and the availability of SSDs and newer energy saving modes, this paper argued for a renewed attention towards improving storage energy consumption. More importantly, we showed that achieving energy proportionality can have significant benefit above and beyond using individual energy-efficient components. Analysis of enterprise data center workloads indicated 40% reduction in storage energy under ideal storage energy proportionality, while another indicated a potential 75% reduction. Our work also outlined the challenges and analyzed techniques for building energy proportional systems, while aligning the techniques with application requirements and workload characteristics. Based on the techniques we have identified, we believe that energy proportionality is achievable in storage systems at a time scale that will make sense in real world environments.

Currently, we are investigating potential energy savings from real-world enterprise settings, and building a simulation infrastructure to experimentally demonstrate the utility of the techniques we have identified. We believe that storage energy management must be coupled into data center energy management, and energy proportionality helps realize that goal. Without such a integral approach, energy management will not be able to approach optimal conservation.

References

[1] Intel Pentium 4 Processor on 90 nm Process, 2005. `http://download.intel.com/design/Pentium4/datashts/3056103.pdf`.

[2] Intel Xeon Processor 5500 Series, 2009. `http://www.intel.com/Assets/PDF/Prodbrief/xeon-5500.pdf`.

[3] Is storage top energy hog in data centers? `http://searchstorage.techtarget.com/news/article/0,289142,sid5_gci1285060,00.html`.

[4] M. Allalouf, Y. Arbitman, M. Factor, R. Kat, K. Meth, and D. Naor. Storage Modeling for Power Estimation. In *ACM SYSTOR*, 2009.

[5] L. A. Barroso and U. Hlzle. The Case for Energy-Proportional Computing. *IEEE Computer*, 40(12), 2007.

[6] T. Burd, T. Pering, A. Stratakos, and R. Brodersen. A Dynamic Voltage-Scaled Microprocessor System. In *IEEE ISSCC*, 2000.

[7] D. Colarelli and D. Grunwald. Massive Arrays of Idle Disks for Storage Archives. In *SC*, 2002.

[8] S. Gurumurthiy, A. Sivasubramaniamy, M. Kandemiry, and H. Frankez. DRPM: dynamic speed control for power management in server class disks. In *13^{th} ISCA*, 2003.

[9] D. Narayanan, A. Donnelly, and A. Rowstron. Write Off-Loading: Practical Power Management for Enterprise Storage. In *6^{th} FAST*, 2008.

[10] D. Narayanan, A. Donnelly, E. Thereska, S. Elnikety, and A. Rowstron. Everest: Scaling down peak loads through I/O off-loading. In *8^{th} FAST*, 2008.

[11] A. E. Papathanasiou and M. L. Scott. Energy Efficient Prefetching and Caching. In *USENIX*, 2004.

[12] E. Pinheiro and R. Bianchini. Energy conservation techniques for disk array-based servers. In *18^{th} ICS*, 2004.

[13] D. Reinsel. The Real Costs to Power and Cool All the World's External Storage. In *IDC Report, Doc 212714*, 2008.

[14] G. Schulz. Storage Industry Trends and IT Infrastructure Resource Management (IRM), 2007. `http://www.storageio.com/DownloadItems/CMG/MSP_CMG_May03_2007.pdf`.

[15] Seagate Technology. Seagate's Sound Barrier Technology (SBT). `http://www.seagate.com/docs/pdf/whitepaper/sound_barrier.pdf`.

[16] N. Tolia, Z. Wang, M. Marwah, C. Bash, P. Ranganathan, and X. Zhu. Delivering Energy Proportionality with Non Energy-Proportional Systems—Optimizing the Ensemble. In *USENIX HotPower*, 2008.

[17] C. Weddle, M. Oldham, J. Qian, A.-I. A. Wang, P. Reiher, and G. Kuenning. PARAID: A Gear-Shifting Power-Aware RAID. In *5^{th} FAST*, 2007.

[18] Q. Zhu, F. M. David, C. F. Devaraj, Z. Li, Y. Zhou, and P. Cao. Reducing Energy Consumption of Disk Storage Using Power-Aware Cache Management. In *10^{th} HPCA*, 2004.

[19] Q. Zhu, Z. Chen, L. Tan, Y. Zhou, K. Keeton, and J. Wilkes. Hibernator: helping disk arrays sleep through the winter. In *20^{th} ACM SOSP*, volume 39, 2005.