Integrated Approach To Data Center Power Management

Lakshmi Ganesh, Hakim Weatherspoon, Tudor Marian, Ken Birman
Computer Science Department, Cornell University
{lakshmi,hweather,tudorm,ken}@cs.cornell.edu

Abstract—Energy accounts for a significant fraction of the operational costs of a data center, and data center operators are increasingly interested in moving towards low-power designs. Two distinct approaches have emerged towards achieving this end: the *power-proportional* approach focuses on reducing disk and server power consumption, while the *green data center* approach focuses on reducing power consumed by support-infrastructure like cooling equipment, power distribution units, and power backup equipment. We propose an integrated approach, which combines the benefits of both. Our solution enforces power-proportionality at the granularity of a rack or even an entire containerized data center; thus, we power down not only idle IT equipment, but also their associated support-infrastructure. We show that it is practical today to design data centers to power down idle racks or containers—and in fact, current online service trends strongly enable this model. Finally, we show that our approach combines the energy savings of power-proportional and green data center approaches, while performance remains unaffected.

Index Terms—Cloud, power management, distributed storage

1 Introduction

Global-scale online services typically run on hundreds of thousands of servers spread across dozens of data centers worldwide. Google is estimated to own over a million servers, while Microsoft's Chicago data center alone is estimated to contain over 300,000 servers [13]. These scales will only increase significantly as Infrastructure-, Platform-, and Storage-as-a-Service (IaaS, PaaS, and SaaS) models mature [1], and approach what many perceive as a likely vision of the future—a handful of infrastructure providers hosting the world's data and computation. As companies compete to take the lead in this space, the operational efficiency of their massive data centers assumes central importance.

This paper focuses on a key aspect of data center operational efficiency—energy management. Energy costs can account for over 35% of the total cost of ownership (TCO) of data centers [5, 6]. As servers grow ever more powerful, and data center server densities continue to increase, Wattage per square foot has been growing as well. This compounds the amount of heat generated per square foot, in turn requiring the expenditure of more energy to remove. Energy costs now rival server costs [10], yet average data center energy efficiency is a mere 50% [5, 6].

Given the economic as well as environmental impact of the global data center energy footprint, it is perhaps surprising that average energy efficiency in this sector is so low. Two challenges impede progress in this space: idle resource energy consumption, and support-infrastructure energy consumption. Since data centers are provisioned for peak load, which is significantly higher than average load, average data center resource utilization is very low [8]. This leads to considerable resource idleness on average, and as idle resources can consume almost as much energy as active ones, significant amounts of energy can be wasted here [9]. The approach typically taken to address this problem is to power down idle resources (disks and servers), to achieve *power proportionality*. A power proportional system consumes energy proportional to its load. It is hard, however, to achieve power proportionality without degrading performance, as resource idleness is difficult to predict accurately.

The second source of data center energy inefficiency is support-infrastructure energy consumption. In addition to servers and IT equipment that are doing directly useful work, data centers contain power distribution, power backup, networking, and cooling infrastructure that enable the IT equipment to function correctly, but do not contribute directly to useful work done. Ideally, total energy consumed by the data center should be a small factor (close to 1) of the energy consumed by IT equipment; this factor is called Power Usage Effectiveness (PUE). In reality, however, data center support-infrastructure consumes energy comparable to the IT equipment, leading to industry average PUE of over 2 [5, 6]. Several *green data center* solutions have been designed to address this problem: direct current (DC) power distribution is advocated for improving power distribution efficiency [4]; battery-backed servers improve on the efficiency of a central UPS power backup solution [2]; free cooling—the practice of using outside air to cool the facility, thus obviating the need for power-hungry chillers—significantly improves cooling efficiency [3]. Most existing facilities, however, cannot take advantage of these solutions without significant engineering overhaul.

This paper takes an integrated approach to data center energy management to simultaneously address idle resource energy consumption, and support-infrastructure energy consumption. We argue for a power management approach that powers down racks or even entire containerized data centers, when idle, thus powering down not only servers, but also their associated power distribution, backup, networking, and cooling equipment. Our evaluation shows that shifting to this model combines the energy savings of the power-proportional as well as the green data center approaches, while not impacting performance. We also show that this shift is practical today at very low deployment cost, and that current data center trends strongly enable it.

The rest of this paper is organized as follows. Section 2 presents related work, and explains its limitations. Section 3 describes our solution and identifies various enabling data center practices. We present evaluation results in section 4, and conclude in section 5.

2 Related Work

We describe the current solution space for data center energy management, under the two broad categories of power-proportional solutions, and green data center solutions. The former category of solutions works by identifying, isolating, and enabling the power-down of idle IT resources through various mechanisms. The latter category of solutions reengineers data centers to improve support-infrastructure energy efficiency, and thus data center PUE. In the next section we show how to integrate these two approaches.

2.1 Power Proportional Solutions

The principle behind power-proportionality is that power should track utilization. More formally, the property states that executing a given job should consume a constant amount of energy, irrespective of how much time it takes. This is possible only if base-line power consumption (power consumed when no job is being executed) is zero. In other words, idle resource power consumption must be zero. The reality, however, is that servers consume almost as much energy when idle or lightly loaded, as when heavily loaded [9]. The problem is exacerbated by the fact that most data centers, being provisioned for peak rather than average load, are very lightly loaded on average [8].

So why don't data center operators just turn off idle resources? Many server components also have the ability to operate in multiple power modes (corresponding to commensurate levels of performance), so that they can be manipulated to consume power proportional to their load, or desired level of performance. However, there are several challenges to this approach. First, switching between power modes takes time, and can lead to degraded performance if load is not accurately predicted and resource power modes matched to it. Most services can tolerate very little, if any, performance degradation. Second, server load is hard to predict accurately. Finally, average server idle times are very short, leading to no energy saving (and perhaps energy wastage) from server power down.

Power proportional solutions address these challenges through various load concentration techniques. The basic insight is that if load can be concentrated on a subset of the data center servers in a predictable manner, then the rest can be powered down to save energy without impacting performance. This solution space can be specified using two basic parameters:

1) *Load Localization Target:* Power-proportional schemes attempt to localize load to a subset of the system so that the rest can be powered down. The load localization target parameter encodes this concept. For instance, MAID (Massive Array of Idle Disks) [28] concentrates popular data on a new set of "cache" disks, while PDC (Popular Data Concentration) [29] uses a subset of the original disk set to house the popular data. Power-aware caches [30] attempt to house the working set of spun-down disks in the cache, to increase their idle time. Write-offloading [31] is a technique that can layer on top of each of these solutions to temporarily divert write-accesses from spun-down disks to spun-up ones, and so is a scheme to localize *write* accesses. SRCMap [32] is similar to MAID and PDC (and additionally uses write-offloading), but is a more principled version of both. KyotoFS [34] is similar to write-offloading, but uses the log-structured file system to achieve write diversions.
2) *Architecture:* Power-proportional systems often add levels to the storage hierarchy in order to create resource power-down opportunities. The architecture parameter encodes the storage hierarchy of a given solution. For instance, the standard storage hierarchy puts primary memory (RAM) ahead of spinning disks. Power-proportional storage solutions add spun-down disks to the tail of this hierarchy. MAID uses an additional set of disks (cache-disks) between memory and the original disk set. PDC, power-aware caching, SRCMap, write-offloading, and KyotoFS all use the original disk set, and add no new levels. Hibernator [27] uses multi-speed disks, as does DRPM [35]. HP AutoRAID [41] divides the disk-set into a smaller, high-performance, high-storage-overhead RAID 1 level, and a larger, low-performance, low-cost RAID 5 level. PARAID [42] is a power-aware variant of AutoRAID.

An important shortcoming with all of these solutions is their neglect of the power overheads of power distribution, networking, and cooling. These overheads can account for as much as 35% of the power consumed by the data center [5, 6]. Compare this with disk power, which accounts for less than 27% of facility power [27]. Yet most solutions in the power proportional space focus on disk or server power down. This narrow focus inherently limits the energy saving potential of these solutions.

2.2 Green Data Center Solutions

We describe some engineering solutions designed to reduce support-infrastructure energy consumption. These fall under three broad categories:

1) Power Distribution Efficiency: For every Watt of energy used to power servers, up to 0.9 W can be lost in power distribution [4]. To a large extent, these losses result from the series of alternating current (AC) to direct current (DC), and DC to AC conversions that are part of the power distribution process. For example, power is typically delivered to a data center as high voltage AC power; this is stepped down to lower voltage AC power for distribution to racks for use by servers and other IT equipment. Inside this IT equipment, power supplies typically convert the AC power to the DC power needed for digital electronics. If the facility uses a UPS, an additional level of indirection is injected in routing the power through the UPS - resulting in another set of AC-to-DC, and DC-to-AC conversions. Power is lost at each of these conversions; further, more power is needed to cool the conversion equipment [4].

   It has been shown that power conversion efficiency can be improved significantly if the data center is supplied with DC power instead of AC power. DC power delivery systems have been shown to be up to 20% more efficient that AC delivery [4]. This solution is orthogonal to ours, and can be used in conjunction with it.

2) Power Backup Efficiency: In order to prevent outages, data centers use a backup power supply that can kick in temporarily if the primary supply fails. Traditionally, this backup takes the form of a central UPS; power to the facility flows through the UPS, charging it, and is then routed to the racks. Significant power loss can result from this model, as the average UPS has an efficiency of only about 92% [2].

   A new model has been demonstrated by Google [2], where power backup is provided through per-server batteries; this distributed design has been shown to achieve close to 100% efficiency [2]. Again, this solution is complementary to ours.

3) Cooling Efficiency: An industry rule-of-thumb suggests that for every Watt of energy consumed by a server, about 0.5 W is needed to remove the resulting heat generated [3]. Data center cooling infrastructure typically consists of a chiller unit to chill the coolant used (water or air), and fans to direct cool air towards the servers, and hot air away from the servers. These are both thermodynamically complex and power-hungry processes.

   A highly effective way to reduce cooling energy consumption is through free cooling, a system that uses ambient air for facility cooling, thus obviating the need for power-hungry chillers. It has been shown that free cooling can help bring data center PUE down to as low as 1.07 [15]. However, existing data centers would need a significant engineering overhaul to adopt this solution. Further, a limiting factor for this solution is the requirement that ambient temperatures be suitable for use in facility cooling.

We now describe a solution with very low deployment overhead, that combines the benefits of the power proportional and green data center approaches.

3 Integrated Approach

As we saw in section 2, current data center energy management solutions are siloed into two separate approaches: power-down solutions for idle IT resources, and engineering solutions for reducing support-infrastructure energy consumption. In this section, we show how to integrate these two approaches by extending power-down solutions to include support-infrastructure.

3.1 Larger Power Cycle Units

We define the power cycle unit (PCU) as the resource unit that the power management scheme operates over. This is the unit whose power state is manipulated to track utilization. For example, disk power management schemes manipulate the disk power state (ON/OFF/possibly low-power states corresponding to lower speeds); CPU power management schemes manipulate CPU power (typically through frequency tuning). Our contention in this paper is that larger PCU options, which have not been explored thus far, promise significantly bigger energy savings.

Figure 1 illustrates our rack PCU model. Depending on the rack and server dimensions, a rack could contain anywhere between 10 to 80 servers, or more. In Figure 1, we show a module consisting of two racks, which share an in-rack cooling system, a rack power distribution unit (PDU), and a top-of-rack switch. The in-rack cooling system [40] draws hot air from the servers in the racks, and circulates cool air to maintain the required server operating temperature. This cooling system would typically be allied with a central chiller unit, which would supply it with chilled air; if the outside air conditions are favourable, the chiller can be dispensed with in favor of free cooling. The rack PDU supplies power to the

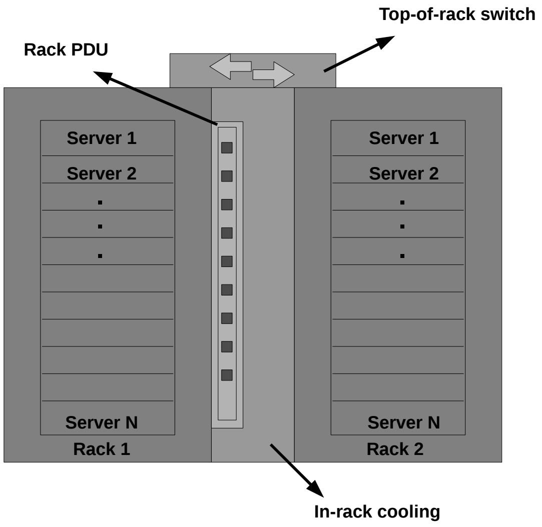
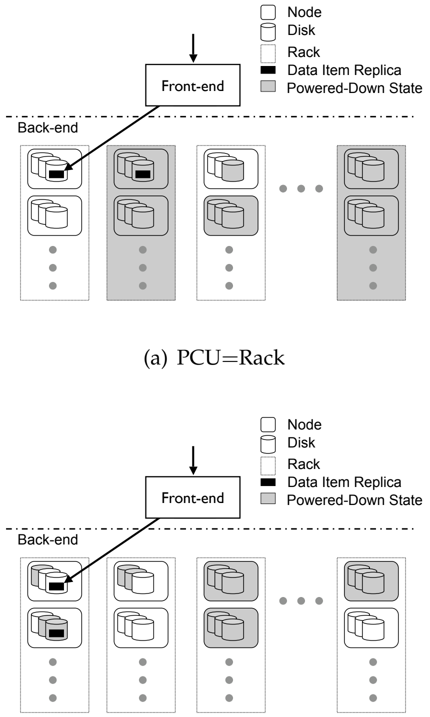


Fig. 1. Rack Power Cycle Unit



(a) PCU=Rack

(b) PCU=Node

Fig. 2. System Model

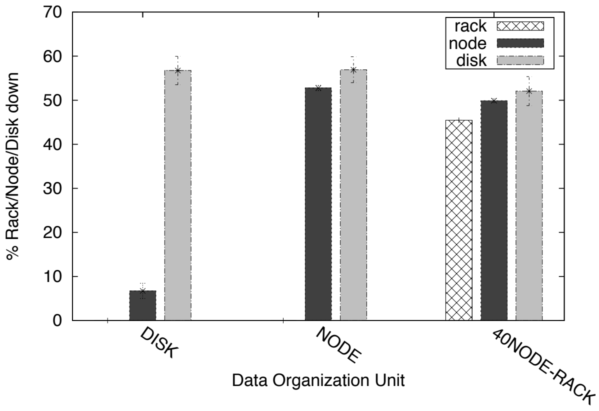


Fig. 3. Impact of Data Organization Scheme on PCU Power-Down Opportunities

rack components; a switched PDU [39] will allow remote control of this power supply, allowing the rack to be turned on or off over the network. Finally, the top-of-rack switch connects the servers in the rack to the data center network. The switch power is also controlled by the rack PDU. The data center network is typically hierarchical, with rack switches connected using row switches, which in turn connect to a set of central switches that have a link to the outside. In this model, the rack PCU can be powered down/up without impacting the rest of the data center network. The breakdown of power draw within the rack PCU depends on a number of factors, such as server power ratings, disk power ratings, number of disks per server, and chilling technique employed (free air cooling/chiller unit). For what the Green Grid characterizes as the average data center [49], servers draw only about 30% of the rack power, while up to 45% may go towards cooling, and the remaining power is spent on the PDU, the switch gear, and power backup.

While racks today are physically self-sufficient, and offer fault isolation from the rest of the data center network, powering them down can result in data unavailability or service interruption unless mindful load placement is practised. In order to create rack power-down opportunities, *PCU-aware data organization* must be employed, as follows:

1) Each data item must be spread (striped/mirrored) *across* PCUs, rather than within them. Thus, assuming some degree of data redundancy, one or more host PCUs may be powered down without impacting the availability of that item.
2) Data access must be localized (as far as possible) to a subset of the PCUs so that others idle, and may be powered down. For read accesses, this is achieved by serving the request from a replica that is on a powered-up disk. For write accesses, this is achieved through write-offloading [31]; if there is a write access to a file, and one or more replicas of that file are on disks that are powered down, then the update must be temporarily diverted to reserve storage on powered-up disks. Periodic cleaning returns these temporary replicas to their original locations. If there is an access (read or write) to a file, all of whose replicas are on powered-down disks, then the access has to suffer the significant performance penalty of waiting for a replica to be powered up. The probability of this should be low by design.

Figures 2(a), and 2(b) illustrate PCU = Rack, and

PCU = Node, respectively. Note how replica placement changes with PCU; also, the creation of idle PCUs through selective access of more active replicas. Figure 3 demonstrates the importance of PCU-aware data organization. We simulate a production data center, and set the PCU to 40-node racks; we then vary the data organization unit (the unit across which replicas are distributed). Notice that unless replicas are distributed across the given PCU (40-node racks, in this case), there is no opportunity for powering them down. Thus, PCU-aware data organization (and retrieval) is key to enabling larger PCUs.

3.2 Enabling Trends

Before evaluating the benefits of shifting to larger PCUs, we discuss the overhead of deploying such a solution today. One of the strengths of our approach is that it is facilitated by several current trends in large-scale online services, thus making the expected deployment overhead quite low.

Rack-and-Roll: Rapid scalability is an imperative for hosting successful online services, and this has led to the so-called rack-and-roll data center expansion model [38]. Data center operators can rapidly expand their facilities by purchasing “commodity racks”, which have servers, top-of-rack switches [37], power distribution units [39], and in-rack cooling equipment [40] pre-installed. Purchasing and commissioning a rack is now a mere matter of hours. Thus, our model rack of Figure 1 is a widely prevalent reality today. Further along this path, entire data centers have now been commoditized—the data center shipping container.

Data Model and Placement: Industry-leading storage designs are converging on certain techniques for performance and reliability that prove strongly enabling for power management solutions in general, and large PCUs in particular:

- *Replication:* Most large-scale systems today replicate their data for fault-tolerance. A replication factor of three is an industry standard [51, 53, 54]. With appropriate replica placement, there is opportunity for powering down one or more replica hosts, without impacting data availability. This provides a tunable parameter—number of live replicas—which can be adjusted based on load, and is a key enabler for storage power management. When combined with PCU-aware replica placement (see trend below), larger PCUs are facilitated.
- *Cross-failure-domain replica placement:* Each object is replicated, not only across disks, but across racks, and also across data centers. This ensures data availability in the face of domain-correlated failures, such as a rack or data center outage. This practice has been adopted in leading systems like Amazon S3 [11], Microsoft Azure [12], and Google GFS [51]. Thus, the mechanism is already in place to support PCU-aware data placement.
- *Append-only model:* A data model that is gaining popularity today due to its performance properties is one where data is stored on disk in immutable data structures. Updates become appends in this model, and consolidation happens lazily. This model caters especially to workloads that are dominated by new writes and large sequential reads, with updates being relatively infrequent. GFS [51], Bigtable [52], and Cassandra [54] are industry-leading systems that use this model. This model is a good fit for power management—updates do not require powering up of all replicas; instead, they can be ‘offloaded’ (appended) to powered-up disks, and lazily consolidated when the requisite replica hosts are up.

Data and Compute Locality: A challenge in data-intensive compute systems is to localize data and computation. Several techniques have been developed that facilitate this. For example, Bigtable [52] exposes data locality control to its clients, so that they can enforce contiguous placement of related data. Another technique is proposed in GreenHDFS [50], which determines object placement by its age; their measurement of a large Hadoop deployment showed that data popularity is strongly correlated with its age in the system, and by placing data of similar age together, they achieve access locality. Thus, mechanisms are in place today in most production systems to ensure data and compute locality. This facilitates power management, because it allows us to power-manage storage without impacting computation; further, it allows us to power down not just disks, but the associated servers as well—in this model, compute tasks assigned to a server are associated with the data hosted on that server, and thus it is reasonable to infer an idle CPU associated with idle disks.

Data Deluge: Studies suggest that the digital universe—all the digital data being created by consumers and businesses worldwide—is growing at a compound annual rate of 57% [22]. Just for the year 2010, this rate of growth translated to an increase in the world’s digital data by 1.2 million petabytes [17]. This trend is significantly changing storage needs. Our belief is that we have arrived at a point in the data deluge where the fraction of data accessed, or even accessible, for any reasonable length of time (a week, say), is a tiny fraction of the total data stored. We come, therefore, to the workload property that the vast majority of data is seldom accessed, the data that is accessed is accessed mostly as reads, and writes that are performed are mostly new writes, instead of updates. This property is highly conducive to power management—it creates opportunities for a significant fraction of the storage system to be powered down without impacting performance or data availability.

TABLE 1
Simulator Parameters (applicable unless specified otherwise)

Parameter	Description	Value
Data Layout	Redundany scheme employed	PCU-aware, 2-way mirroring
Disk Power (W) (Up/Down/Tran)	Power consumed by disk when up, down, or transitioning between up and down	10/2/10
Node Power (W) (Up/Down/Tran)	Power consumed by node (over and above that consumed by its disks) when up, down, or transitioning between up and down	200/5/200
Rack Power Overhead (%) (Up/Down/Tran)	Power consumed by rack (over and above that consumed by its nodes) when up, down, or transitioning between up and down	50/0/50
Disk Access Time (ms)	Time taken to retrieve 1 byte from disk that is up	8
Disk Bandwidth (MBps)	Data transfer rate from disk that is up	120
Disk Transition Time (s)	Time taken by disk to go between up and down states	6
Node Transition Time (s)	Time taken by node (over and above that taken by its disks) to go between up and down states	30
Rack Transition Time (s)	Time taken by rack (over and above that taken by its component nodes) to go between up and down states	300
Power Check Interval (hr)	The intervals at which all PCUs are examined and idle ones powered down	0.5
Power Management Start Time (hr)	The interval after start of simulation when power checking begins	0.5
Disk Power Down Threshold	An exponentially weighted disk access count threshold below which the disk is considered idle	10
Per-Server Cache Size	self-explanatory	1 GB
Number Of Nodes	Actual number from an IA MC data center	840
Number Of Disks/Node	Actual number from an IA MC data center	4

One concern with any power-down solution is the potential impact of power-cycling on component reliability, and hence on overall system availability. With our approach of using larger PCUs, a natural question to ask is whether powering down cooling equipment increases their failure rates. Experiential evidence suggests that it does not [7]: Economizer modes of operation, which bypass (power down) chiller units when environmental conditions allow use of ambient air, are already extent in several production facilities, and have not impacted chiller reliability.

In summary, with online service trends leading to huge amounts of replicated, potentially seldom-accessed data, there is increasing opportunity for saving energy in data centers by powering down idle/redundant resources. Further, current industry practices are making racks (and more recently, containerized data centers) the unit of choice for resource commissioning as well as replication, which in turn enables larger PCUs. The deployment of a larger PCU solution would thus incur little overhead, and have large energy-saving potential. Next, we quantify this energy-saving potential.

4 Evaluation

The aim of this study is to quantify the potential energy savings from using larger PCUs, for different data center settings. We address the following questions:

1) How much energy can be saved by shifting to larger PCUs?
2) How is performance impacted by shifting to larger PCUs?
3) Under what conditions does it make sense to shift to larger PCUs?

We describe our methodology, and then present our findings.

4.1 Methodology

We use simulations to explore a number of different data center settings and PCU options. Our results have been promising, and we are planning a small-scale implementation study in future work.

4.1.1 Simulator

We model a large distributed storage system comprised of a set of servers and their disks, and associated support-infrastructure. The model allows PCU choices of disk, node, rack (comprising a specified number of nodes as well as their cooling and power distribution equipment), and container (comprising a specified number of nodes, their cooling and power distribution equipment, as well as a UPS). Parameters that can be varied include: disk/server/rack/container power ratings, number of disks per server, number of servers per rack/container, per-server memory capacity, per-disk storage capacity,

TABLE 2
Trace Characteristics

Attribute	Trace 1	Trace 2	Trace 3
Duration	6 hrs	6 hrs	6 hrs
# accesses	6.5m	7m	6.6m
Avg. access size (MB)	1.7	1.3	1.5
Max access size (GB)	7.73	20.74	7.73
Avg # accesses to a node	7797.77	8338.12	7862.95
Max # accesses to a node	110322	184424	120983
# Nodes accessed	833	838	835

data organization scheme (PCU-aware or not), data replication scheme (full replication, striping, RAID schemes), power management policy (idleness threshold for powering down resource units, how frequently power management decisions are made), among others. Given the system specifications, we simulate the progress of each file request through the system, recording latency and power consumption. Periodically, all the disks in the system are checked to see whether their access frequency is less than a given idleness threshold; if so, they are powered down. If the system PCU is set to node or bigger, then nodes with all their disks down and with a cache access frequency lower than the specified threshold, are powered down as well. Finally, for systems with PCU larger than a node, a PCU is powered down if all its nodes are down. An access to a powered-down disk will trigger the powering up of that disk (including, if needed, the powering up of its parent node and PCU). Table 1 presents the standard simulation settings.

4.1.2 Data

To drive our simulations, we use access logs from a production data center hosting the Internet Archive's (IA) Media Collection [36] service. This is a web service that serves text, image, audio, and video files from a large (2 PB) collection. Table 2 gives details of these traces. Unless otherwise specified, each data point presented in the following section is the averaged result of running 6-hour traces from three different days of the week of April 3-9, 2009. (a Monday, Tuesday, and Friday, the same set of hours being picked from each day). The traces are HTTP GET logs, and specify, for each file access, the access time, the file name, and file size, as well as the target server ID and disk ID. However, we manipulate this information slightly to conform to different data organization layouts. Given a data organization scheme—PCU-aware, 2-way mirroring, for example—we statically map each disk to a "mirror disk" such that the mirror disk is on a different PCU from the original disk. An access request to any item on either disk is then directed to the more active of the two. Support for dynamic, per-file mapping is planned in future work.

The traces have a read-ratio ($\frac{\text{\# reads}}{\text{\# accesses}}$) of very close to 1 (0.9926), and so we did not model writes (we ignored writes on the trace). We argue that this lack of writes does not bias our results unduly: From a power management point of view, write accesses (with write-offloading) are indistinguishable from read accesses that miss the cache; ie., their pattern of disk access will be similar. Our traces, having been captured at the backend, are comprised entirely of cache misses from the frontend web server caches, and thus have inherently low cacheability. This is reflected in our simulations, where they show less than 54% combined hit rate at the server caches. Thus, we expect that the disk access patterns would not change significantly even if up to 40% of our accesses were writes (as they would behave similarly to the 46% cache misses we currently observe). There is one exception to this rule: some resource power ups may be triggered when reclamation of offloaded writes is done. Since this is an infrequent event, its impact on energy consumption should not be significant. In future work, however, we plan to add write-offloading support to our simulator and verify this intuition.

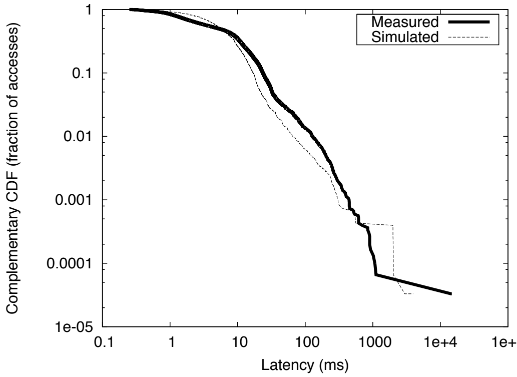


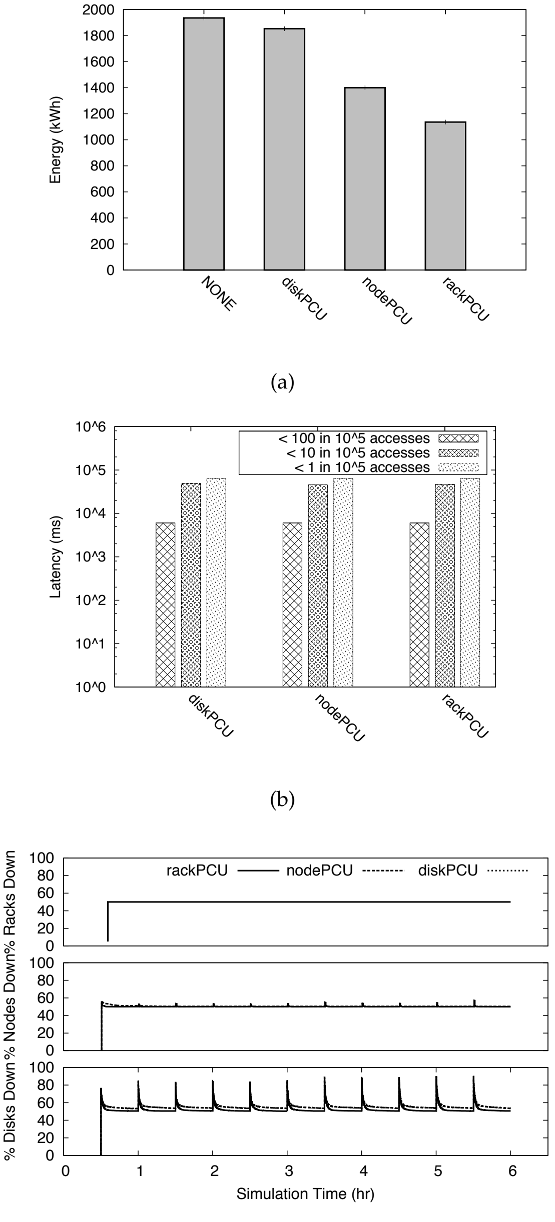
Fig. 4. Comparing the Simulator Against Single-Server RAID-0 Setup

4.1.3 Validation

We used two methods to ensure that our simulator tracks ground truth. First, we compared its storage model at the granularity of a single server against measurements from a real storage node. Second, we used actual measurements from production settings to configure the simulator's rack-level parameters.

Figure 4 compares measured and simulated access latencies from a single server RAID-0 store spread over six disks. Each data block is striped over the six disks (no redundancy). We combined file access traces from three of the most-accessed nodes in the IA data, and replayed them on the RAID-0 system. This combined trace spanned 25 minutes, and comprised 32,749 requests. We also ran this trace on our simulator, configured to resemble the RAID-0 setup. As seen in Figure 4, our simulator tracks reality well at the server level.

We obtained node and rack power cycling information from actual measurements at the IA. These have informed our choice of node and rack transition times, and power overheads. In future work, we plan to verify our



(a)

(b)

(c)

Fig. 5. Optimal PCU Size for the Internet Archive

simulator at rack-level through a small-scale prototype implementation.

4.2 Results

We now explore the potential as well as the limitations of power management through larger PCUs. We present our results in the context of three motivating scenarios:

4.2.1 Motivating Scenario: Online Media Service

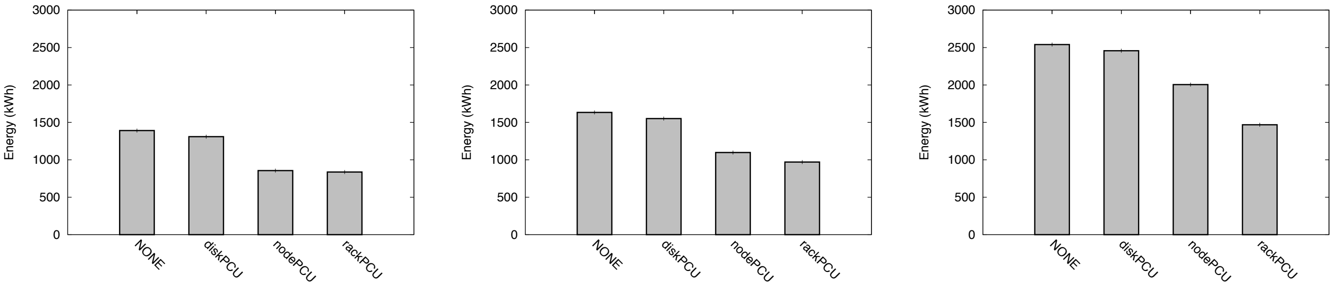
The Internet Archive's Media Collection (MC), which stores and serves over 2 PB of video, audio, image and text files, is a popular online media service. Our workload data derives from one of the MC data centers; we now explore the right choice of PCU for this data center.

We simulate an MC data center; Table 1 describes the configuration parameters, which are intended to reflect ground truth. The IA maintains two copies of each file, on two separate storage servers. The MC data center we simulate has 840 storage servers—commodity machines with 4 disks each. We compare PCU choices of disk, node, and rack. We configure rack overhead to be 50%; ie., the support-infrastructure on each rack consumes 50% as much power as the servers and disks on the rack. We assume that an additional 10% power goes to the data center power backup equipment. We show results for a 42-server rack, but we confirmed that very similar results obtain for a whole range of rack sizes (from 10-server to 200-server), so long as rack power overhead and transition time are constant.

Figure 5(a) shows that rack PCUs lead to 41% energy savings (a 9.7X improvement over disk PCU, and 1.5X improvement over node PCUs) in this data center. Further, we see in Figure 5(b) that the system performance under the rack PCU configuration is the same as its performance under the node PCU configuration; each set of three bars in this graph shows the highest latency seen in the 99.9-, 99.99-, and 99.999- th percentile of accesses respectively (left to right). Figure 5(c) explains why the rack PCU configurations do not impose any performance penalty. For each configuration, it tracks the number of racks, nodes, and disks that are powered down over the length of the simulation. We see that for the rack PCU configuration, the number of racks down stays constant after the initial power check interval. This means that no access goes to a powered-down rack, with the result that rack power-downs have no additional performance penalty.

Figure 6 shows the impact of PUE on optimal PCU size. Rack power overhead reflects data center PUE—in our model, 50% rack power overhead implies a PUE of at least 1.6 (factoring in the additional 10% overhead for the UPS). In other words, when PUE is X, for every Watt consumed by the servers in a rack, 0.1 W goes towards the UPS, and (X-1.1) W goes towards the in-rack cooling unit. We see that for values of PUE below 1.35, larger PCUs no longer make sense—it is better to use node-based power management in these settings. This bears out our intuition—the motivation for shifting to larger PCUs is to reduce some of the non-IT power overheads of the data center; the smaller these overheads, the less reason to make this shift. Keep in mind, however, that the industry average for data center PUE is 2-2.5.

Figure 7 shows the impact of disk-to-CPU ratio on optimal PCU size. For a service such as the IA Media Collection, whose load is entirely I/O-bound, it makes sense to use servers with a larger number of disks. This is in fact precisely the direction the IA is taking; they are in the process of transitioning to storage nodes with 24 to 36 disks each. In this disk-heavy model, we reexamine optimal PCU choice. Note that, when maintaining the same data center capacity and increasing the disk-to-node ratio, the number of nodes (and racks) decreases.

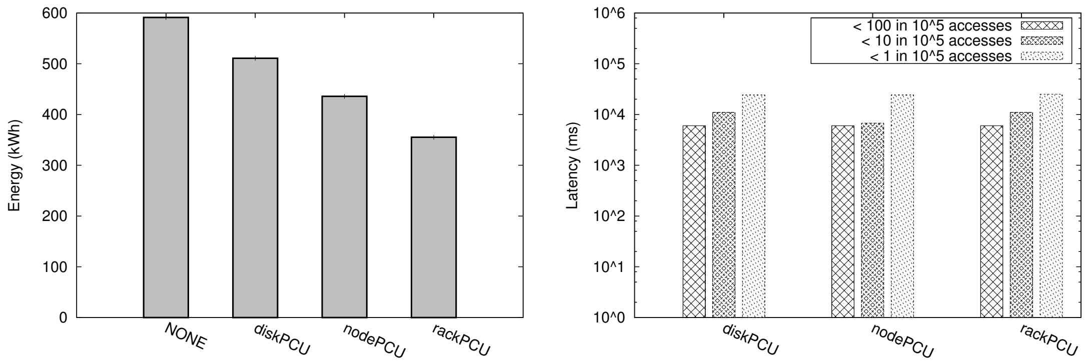


(a) PUE=1.15

(b) PUE=1.35

(c) PUE=2.1

Fig. 6. Effect of PUE on Optimal PCU Size



(a)

(b)

Fig. 7. Optimal PCU Size When Disk-to-CPU Ratio is 24

Figure 7 shows that a rack is still the optimal PCU choice when disk-to-CPU ratio is increased to 24; comparing with Figure 5 (disk-to-CPU ratio of 4), however, as might be expected, we see that the energy savings over disk-based power management has decreased.

4.2.2 Motivating Scenario: Storage as a Service

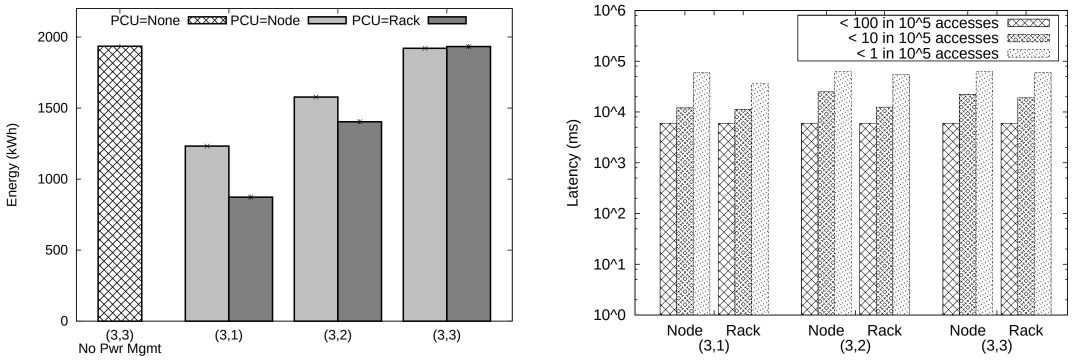
We now consider another popular online service—Storage as a Service (SaaS). Amazon's Simple Storage Service (S3) [11], for example, provides storage at approximately 10 cents per Gigabit-month. SaaS providers typically replicate data for reliability - the basic service providing at least 3-way replication, with replicas spread across failure domains such as racks and data centers. Clients can alternatively choose a cheaper option—lower level of replication for data requiring less stringent reliability guarantees.

Consider a new SaaS feature: tunable number of live replicas. Clients, when they upload objects, can specify their expected popularity, and tune the number of replicas that need be kept live; the lower this number, the lower the cost of storing the object. With mechanisms already extant for spreading replicas across racks (and data centers), PCU-aware data organization is an easy fit. Figure 8 shows the energy savings from reducing the number of live copies. The number of live replicas is represented as (r,l) along the x-axis, where r is the total number of replicas (3, here), and l is the number of live ones. We show results for two choices of PCU: node, and 40-node rack. We see that keeping only one copy live in the rack PCU configuration leads to 55% energy savings, while keeping two copies live saves 27% energy. Assuming that energy costs contribute 30% to total storage cost, these savings could reduce end-user perceived storage prices by a significant 16.5%, or 8% respectively.

4.2.3 Motivating Scenario: Container Farm

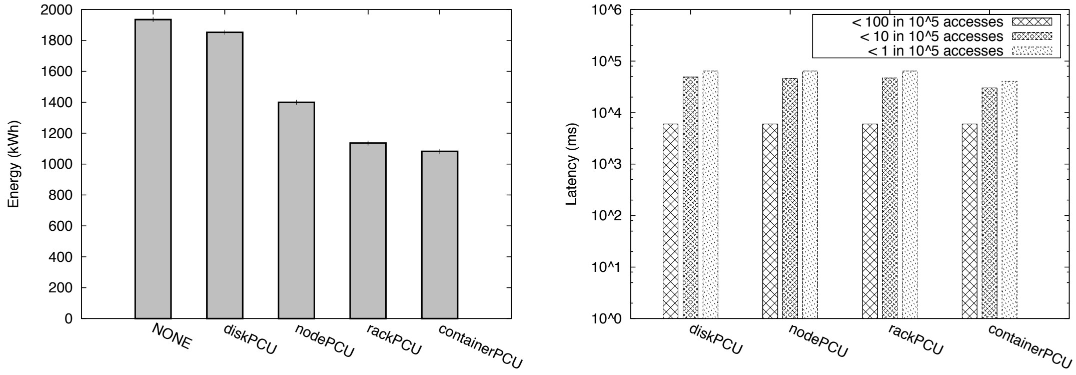
Containerized data centers are seeing increasing adoption in industry; for example, Microsoft reportedly owns a facility in Chicago comprising 112 containers—a container farm [14]. Containers have the advantages of modularity, ease of deployment, ease of management, and improved space and power efficiency, and might reasonably be expected to be a popular data center commissioning unit of the future. With this in mind, we consider the right PCU choice for a data center consisting of a network of containers.

In this model, we have a new PCU choice—an entire container. The advantage of powering down a container is that we power down its associated power distribution and backup infrastructure. Assuming that these overheads add up to 10% of the power draw, Figure 9 shows the energy savings from container-based power



(a) (b)

Fig. 8. Energy Savings From Tuning Number of Live Replicas



(a) (b)

Fig. 9. Optimal PCU Choice for a Container Farm

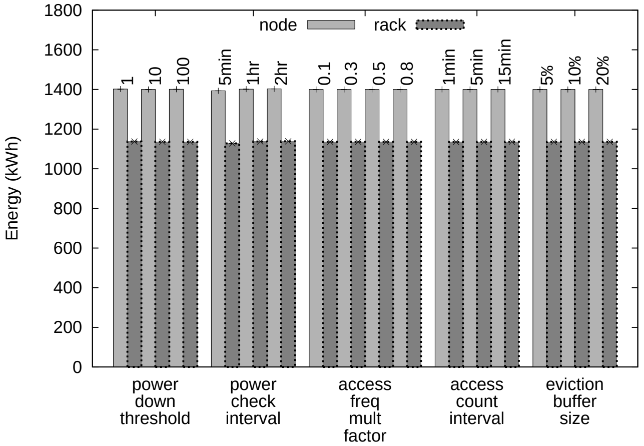


Fig. 10. Result Sensitivity to Simulator Settings

management. Here, our data center consists of two 420-node containers, container power-up takes 10 minutes (as opposed to 5 minutes for rack power-up), and container power overhead is 60% (as opposed to 50% rack power overhead). We see that the container PCU and rack PCU offer similar energy savings (Figure 9(a)), and similar performance (Figure 9(b)).

4.3 Sensitivity Analysis

Finally, we verify that our results are not artifacts of the simulator settings. Figure 10 shows that our findings are robust to simulator fine-tuning. *Disk power down threshold* is the access frequency threshold below which a disk is considered idle (and hence can be powered down). Periodically (with period length = *disk power check interval*), all disks (as well as higher-level components) are checked for power-down opportunities. Disk access frequency is computed as an exponentially weighted moving average, with a *multiplicative factor* determining the weight given to the mean frequency computed over the most recent interval; *access count interval* is the length of this interval. Finally, *cache eviction buffer* allows multiple evicted cache entries to be aggregated before being evicted together; its size is measured as a percentage of cache size. As Figure 10 shows, none of these simulator-specific parameters plays any role in determining the simulation results.

4.4 Summary

To summarize, we have examined a number of different online service models and shown that in each case significant energy savings are achieved by use of larger PCUs. In the Internet Archive setting, we have shown that shifting to rack PCUs achieves 9.7X more energy savings than disk-based power management. This translates to a saving of about 2.9MWh per day over that of a disk-based solution, even for a small 840-node facility. However, we note that the benefit of larger PCUs is strongly tied to the facility PUE—if PUE falls below 1.35, larger PCUs are no longer optimal.

We believe that an increasingly likely vision of the future of online services is one where a few infrastructure providers compete to host the world's services and data. We show that for an SaaS provider, existing data replication and placement policies fit our large PCU model. Further, we show that an SaaS provider could provide storage options up to 16.5% cheaper by adopting rack-based power management, and tuning the number of replicas kept live.

Finally, we examine another point in the design space—container farms. We show that, in this scenario, using entire containers as the PCU is practical, and leads to no performance penalty over node-based power management.

5 Conclusion

With online services continuing to grow in size and number, the power efficiency of the data centers that host them has assumed central importance. Current power proportional designs limit themselves by focusing only on the power consumed by the IT equipment, and neglecting the significant power draw of non-IT equipment like coolers, power distribution units (PDUs), and power backup equipment (UPSes). In this paper, we show how to take an integrated approach to address these overheads by shifting to larger units of power management—racks, or even entire containerized data centers. We show that such a shift is practical (achievable without impacting performance), simple (much of the mechanism needed to support larger PCUs—cross-PCU replica placement, etc—is already in place in most production systems), and highly beneficial (leading to several times more energy savings than current solutions).

Acknowledgement

We would like to thank Brewster Kahle and the IA folk for sharing their data, and for their whole-hearted support and enthusiasm. We hope this work will be useful to them. We are also grateful to AFRL, NSF, NetApp, and Microsoft for their support of this effort.

References

[1] "The NIST Definition of Cloud Computing", The National Institute of Standards, Technology (NIST), 2009.

[2] Katie Fehrenbacher, "A Key to Google's Data Center Efficiency: One Backup Battery Per Server", GigaOm, April 1st, 2009. http://gigaom.com/cleantech/a-key-to-googles-data-center-efficiency-one-backup-battery-per-server/

[3] Simon Hancock, "Iceland Looks to Serve the World", BBC Click, October 9th, 2009.

[4] My Ton, Brian Fortenbery, Willian Tschudi, "DC Power for Improved Data Center Efficiency", Report by Laurence Berkeley National Laboratory (LBNL), 2008.

[5] Kushagra Vaid, "Datacenter Power Efficiency: Separating Fact From Fiction", Invited Talk at Usenix HotPower, 2010.

[6] Dileep Bhandarkar, Kushagra Vaid, "Rightsizing Servers to Achieve Cost, Power Savings", Microsoft Global Foundation Services Whitepaper, 2009.

[7] John Niemann, John Bean, Victor Avelar, "Economizer Modes of Data Center Cooling Systems", Schneider Electric Data Center Science Center Whitepaper, 2011. http://www.apcmedia.com/salestools/JNIN-7RMNRX_R0_EN.pdf

[8] James Kaplan, William Forrest, Noah Kindler, "Revolutionizing Data Center Energy Efficiency", Report by McKinsey&Company, July 2008. http://www.mckinsey.com/clientservice/bto/pointofview/pdf/Revolutionizing_Data_Center_Efficiency.pdf

[9] David Meisner, Brian Gold, Thomas Wenisch, "PowerNap: Eliminating Server Idle Power", ACM International Conference on Architectural Support for Programming Languages, Operating Systems (ASPLOS), 2009.

[10] John Timmer, "Datacenter Energy Costs Outpacing Hardware Prices", Ars Technica, 2009. http://arstechnica.com/business/news/2009/10/datacenter-costs-outpacing-hardware-prices.ars

[11] Amazon Simple Storage Service (S3). http://aws.amazon.com/s3

[12] Storing Data in Windows Azure. http://www.windowsazure.com/en-us/develop/net/fundamentals/cloud-storage/#storing

[13] Rich Miller, "Who Has The Most Web Servers?", Data Center Knowledge, May 14, 2009. http://www.datacenterknowledge.com/archives/2009/05/14/whos-got-the-most-web-servers/

[14] Rich Miller, "Microsoft's Windows Azure Cloud Container", Data Center Knowledge, November 18, 2009. http://www.datacenterknowledge.com/archives/2009/11/18/microsofts-windows-azure-cloud-container/

[15] Open Compute Project. http://www.opencompute.org

[16] Rich Miller, "How A Good PUE Can Save 10 Megawatts", Data Center Knowledge, September 13, 2010. http://www.datacenterknowledge.com/archives/2010/09/13/how-a-good-pue-can-save-

10-megawatts/

[17] Bernard Golden, “Cloud Computing: How Big is Big Data? IDC’s Answer”, CIO, May 7, 2010. http://www.cio.com/article/593039/Cloud_Computing_How_Big_is_Big_Data_IDC_s_Answer

[18] Radhika Kaushik, “Spreading the digital word”, ExtremeTech, April 29, 2003. http://www.extremetech.com/article2/0,3973,1047454,00.asp

[19] Emma Woollacott, “Digital content doubles every 18 months”, TG Daily, May 19, 2009. http://www.tgdaily.com/hardware-features/42499-digital-content-doubles-every-18-months

[20] Electric Power Monthly, U.S. Energy Information Administration, January 14, 2011. http://www.eia.doe.gov/electricity/epm/tables5_6_a.html

[21] “Data, data everywhere”, The Economist Special Report, February 25, 2010.

[22] Peter Lyman, Hal Varian, Peter Charles, Nathan Good, Laheem Jordan, Joyojeet Pal, “How Much Information? Executive Summary”, School of Information Management Systems, UC-Berkeley, 2003.

[23] Albert G. Greenberg, James R. Hamilton, David A. Maltz, Parveen Patel, “The Cost of a Cloud: Research Problems in Data Center Networks”, Computer Communication Review, 2009.

[24] Hrishikesh Amur, James Cipar, Varun Gupta, Gregory Ganger, Michael Kozuch, Karsten Schwan, “Robust, Flexible Power-proportional Storage”, Symposium on Cloud Computing (SOCC), 2010.

[25] David Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, Vijay Vasudevan, “FAWN: A Fast Array of Wimpy Nodes”, Symposium on Operating Systems Principles (SOSP), 2009.

[26] Adrian Caulfield, Laura Grupp, Steven Swanson, “Gordon: Using Flash Memory to Build Fast, Power-efficient Clusters for Data-intensive Applications”, Architectural Support for Programming Languages, Operating Systems (ASPLOS), 2009.

[27] Qingbo Zhu, Zhifeng Chen, Lin Tan, Yuanyuan Zhou, “Hibernator: Helping Disk Arrays Sleep Through The Winter” ,Symposium on Operating Systems Principles (SOSP), 2005.

[28] Dennis Colarelli, Dirk Grunwald, Michael Neufeld, “The Case for Massive Arrays of Idle Disks (MAID)”, File and Storage Technologies (FAST), 2002.

[29] Eduardo Pinheiro, Ricardo Bianchini, “Energy Conservation Techniques for Disk Array-Based Servers”, International Conference on Supercomputing (ICS), 2004.

[30] Qingbo Zhu Francis, Francis M. David, Christo F. Devaraj, Zhenmin Li, Yuanyuan Zhou, Pei Cao, “Reducing Energy Consumption of Disk Storage Using Power-Aware Cache Management”, Symposium on High-Performance Computer Architecture (HPCA), Febuary 2004.

[31] Dushyanth Narayanan, Austin Donnelly, “Write Off-Loading: Practical Power Management for Enterprise Storage”, File, Storage Technologies (FAST), 2008.

[32] Akshat Verma, Ricardo Koller, Luis Useche, Raju Rangaswami, “Energy Proportional Storage Using Dynamic Consolidation”, File, Storage Technologies (FAST), 2010.

[33] Lakshmi Ganesh, Hakim Weatherspoon, Ken Birman, “Beyond Power Proportionality: Designing Power-Lean Cloud Storage”, IEEE Network Computing, Applications (NCA), 2011.

[34] Lakshmi Ganesh, Hakim Weatherspoon, Mahesh Balakrishnan, Ken Birman, “Optimizing Power Consumption in Large Scale Storage Systems”, HotOS, 2007.

[35] Sudhanva Gurumurthi, Anand Sivasubramaniam, Mahmut Kandemir, Hubertus Franke, “DRPM: Dynamic Speed Control for Power Management in Server Class Disks”, International Symposium on Computing Architecture (ISCA), 2003.

[36] The Internet Archive. http://www.archive.org

[37] Data Center Top-of-Rack Architecture Design, Cisco, 2009. http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-522337.html

[38] Cisco Data Center Infrastructure 2.5 Design Guide, Cisco, 2007. http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html

[39] Switched Rack PDU, APC. http://www.apc.com/products/family/index.cfm?id=70

[40] “High Density In-Rack Cooling Solutions for Server Racks, Computer Rooms, Server Rooms & Data Centers”, 42U. http://www.42u.com/cooling/in-rack-cooling/in-rack-cooling.htm

[41] John Wilkes, Richard Golding, Carl Staelin, Tim Sullivan, “The HP AutoRAID Heirarchical Storage System”, ACM Transactions on Computer Systems (TOCS), 1996.

[42] Charles Weddle, Mathew Oldham, Jin Qian, An-I Andy Wang, Peter Reiher, Geoff Kuenning, “PARAID: A Gear-Shifting Power-Aware RAID”, File and Storage Technologies (FAST), 2007.

[43] Elliot Jaffe, Scott Kirkpatrick, “Architecture of the Internet Archive”, Israeli Experimental Systems Conference (SYSTOR), 2009.

[44] Bruce Baumgart, Matt Laue, “Petabyte Box for Internet Archive”, November 2003.

[45] Cade Metz, “Sun packs 150 billion web pages into meat locker”, The Register, March 2009. http://www.theregister.co.uk/2009/03/25/new_internet_archive_data_center/

[46] Brewster Kahle, “Project Greenbox”, January 2008. http://backyardfamilyfarm.wikispaces.com/Project+Greenbox

[47] In Personal Communication with Brewster Kahle

and the Internet Archive Staff, January 14, 2010.

[48] The Green Grid. http://www.thegreengrid.org

[49] "Seven Strategies To Improve Data Center Cooling Efficiency", The Green Grid, 2008. http://www.thegreengrid.org/en/Global/Content/white-papers/Seven-Strategies-to-Cooling

[50] Rini Kaushik, Milind Bhandarkar, Klara Nahrstedt, "Evaluation, Analysis of GreenHDFS: A Self-Adaptive, Energy-Conserving Variant of the Hadoop Distributed File System", Cloud Computing Technology, Science (CloudCom), 2010.

[51] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung, "The Google File System", Symposium on Operating Systems Principles (SOSP), 2003.

[52] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson Hsieh, Deborah Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert Gruber, "Bigtable: A Distributed Storage System for Structured Data", Operating Systems Design, Implementation (OSDI), 2006.

[53] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, Werner Vogels, "Dynamo: Amazon's Highly Available Key-Value Store", Symposium on Operating Systems Principles (SOSP), 2007.

[54] Avinash Lakshman, Prashant Malik, "Cassandra - A Decentralized Structured Storage System", Large Scale Distributed Systems, Middleware (LADIS), 2009.

AUTHOR BIOGRAPHIES

Lakshmi Ganesh: Lakshmi received her Ph.D. from Cornell University in January 2012. Her thesis work, under the joint advisorship of Dr. Hakim Weatherspoon and Dr. Ken Birman, was on data center power management. She was the recipient of an MSR Graduate Research Fellowship award in 2009, and a Computing Innovation Fellowship in 2011. She is now working as a postdoctoral fellow with Dr. Mike Dahlin and Dr. Lorenzo Alvisi at the Laboratory for Advanced Systems Research at the University of Texas at Austin.

Hakim Weatherspoon: Hakim Weatherspoon is an assistant professor in the Department of Computer Science at Cornell University. His research interests cover various aspects of fault-tolerance, reliability, security, and performance of large Internet-scale systems such as cloud computing and distributed systems. Professor Weatherspoon received his Ph.D. from University of California at Berkeley and B.S. from University of Washington. He is an Alfred P. Sloan Fellow and recipient of an NSF CAREER award, DARPA Computer Science Study Panel (CSSP), IBM Faculty Award, the NetApp Faculty Fellowship, and the Future Internet Architecture award from the National Science Foundation (NSF).

Tudor Marian: Tudor Marian received his Ph.D. from Cornell University in August 2011. His thesis work, under Prof. Hakim Weatherspoon, was on operating systems abstractions for high-speed software packet processing in datacenter networked environments. After completing his postdoctoral work at Cornell University, he joined Google, where he works with the Storage and Infrastructure group.

Ken Birman: Ken Birman is the N. Rama Rao Professor of Computer Science at Cornell University, where he has headed a research effort in the area of high assurance distributed computing for thirty years. Ken is best known for inventing the virtual synchrony computing model and building the Isis Toolkit, which was ultimately used to build the system that operated the New York Stock Exchange for more than a decade, and the systems that continue to operate the French Air Traffic Control system and the US Navy AEGIS today. He also pioneered in the use of gossip protocols for system monitoring, management and control; several of his solutions are used today in the platforms that operate today's largest cloud computing infrastructures, notably at Amazon, IBM and Microsoft. A Fellow of the ACM since 1999, Ken won the 2009 IEEE Kanai Award for his research in distributed systems.