Temperature Aware Power Allocation: An Optimization Framework and Case Studies

Shen Li, Shiguang Wang, Tarek Abdelzaher
University of Illinois at Urbana Champaign
{shenli3, swang83, zaher}@cs.uiuc.edu

Maria Kihl, and Anders Robertsson
Lund University
Maria.Kihl@eit.lth.se, and andersro@control.lth.se

***Abstract*—In this paper, we propose a general power model along with a versatile optimization methodology that can be applied to different applications for minimizing service delay while satisfying power budget in data centers. We test our methodology on two totally different applications from both cloud computing and enterprise scenarios. Our solution is novel in that it takes into account the dependence of power consumption on temperature, attributed to temperature-induced changes in leakage current and fan speed. While this dependence is well-known, we are the first to consider it in the context of minimizing service delay. Accordingly, we implement our optimization strategies with Hadoop, Tomcat, and MySQL on a 40-node cluster. The experimental results show that our solution can not only limit the power consumption to the power budget but also achieves smaller delay against static solutions and temperature oblivious DVFS solutions.**

***Index Terms*—data center, map-reduce, web-server, thermal-aware optimization, energy management**

I. Introduction

In this paper, we propose a general power model and temperature aware power allocation (TAPA) optimization framework that help to minimize services delay with a given power budget in data centers. We implement and empirically evaluate our framework on a Map-Reduce cluster and multi-tier web servers respectively. The research challenge is to optimize the trade-off between service delay and power consumed. We do so by developing algorithms that minimize delay induced by cluster for any given power budget. Various implementations [7], [8], [9] of Map-Reduce have been developed in recent work. We modify Hadoop in our implementation to embody our optimization solutions. For the web server case, we use Tomcat on the second tier servers, and MySQL on the third tier servers.

Our algorithm is novel in that it accounts for thermally-induced variations in machine power consumption. Prior work on energy-efficiency of Map-Reduce and web server workloads considered power consumption models that are a function of only the energy state and machine load. In reality, both leakage current and fan speed depend on machine temperature. Hence, a higher power consumption is incurred at higher temperatures, even at the same energy state (e.g., DVFS setting) and at the same load. Modifications are thus needed to existing energy management algorithms to account for the aforementioned temperature-dependency of power consumption. While the effect of temperature on leakage current (and, hence, energy) is itself well-known [11], [12], [13], [14], [15], we are the first to consider its implications on performance (namely, latency) and energy trade-offs in data centers.

Our algorithm requires that a power budget be expressed. This can be given by the power supplier. Some data centers do have a fixed peak power budget [17][30] to avoid excessive charges for peak power consumption. Even in scenarios without a global power bound, a local power budget is often used as a proxy for controlling heat distribution since most power consumed by servers turns into heat. Some researchers [23], [22] proposed to divide the whole data center into small zones and apply elaborately calculated power budgets to each zone to counter-balance temperature differences and remove hot spots. Finally, even in systems where no power budget exists at any level, it is useful to optimize the power-performance trade-off. This optimization problem can be generically cast as one of maximizing performance for any given input power. Hence, our algorithm maximizes the computational capacity of the cluster, without exceeding a stated power budget.

Improving the power-performance trade-off may result in significant monetary savings for data center operators. Hamilton *et al.* [1] estimated that, in 2008, money spent on power consumption of servers and cooling units had exceeded 40 percent of total cost for data centers, which reached more than 1 million per month for a single data center. The cost is still increasing as both prices of energy and scale of data centers are on the rise [2]. The U.S. Department of Energy states that, by 2012, the cost of powering up a data center will surpass the original IT investment [3]. Our experiments show that the temperature-dependent component of power contributes a considerable part to the total power consumption of machines. Given the large variations in temperature among different machines (depending on their placement in the cluster) [16], [18], accounting for temperature dependencies is essential to achieving a more efficient cluster energy usage and hence a non-trivial improvement in data center operation costs.

The remainder of this paper is organized as follows. In Section II, we summarize the related works of this paper. Section III demonstrates the general power model, optimization framework and two case studies. Implementation details are described in Section IV. Finally, our experimental results are shown in Section V.

II. Related Work

Current cluster energy saving policies cover several different basic techniques, such as dynamic voltage and frequency scaling (DVFS) [10], [19], [20], [21], [31], workload distribution [24], [25], [29], as well as turning machines on and off [26], [28], [27]. While much work considered energy minimization, other literature [17], [30] focused on maximizing service capacity under a power constraint. Few papers consider temperature as a parameter when designing energy saving policies. Given that most power consumed by servers turns into heat, some algorithms, such as OnePassAnalog [22] , use the amount of power as a proxy of workload, and assign an amount of power to each server inversely proportional to the server's inlet temperature. Moore *et al.* designed the Zone-Based Discretization (ZBD) algorithm [23] to improve the granularity of load and power distribution. When one server consumes more power than it is assigned, this server will "borrow" power from its neighbor servers, *i.e.*, neighbor servers have to use less power than assigned. Banerjee *et al.* [29] proposed Highest Thermostat Setting (HTS) algorithm to smooth out hot spots and increase the set point in CRAC. In their algorithm, each server s_i has its own thermostat setting requirement (TSR), which indicates the highest possible set point of CRAC to satisfy s_i's cooling demand. They rank servers by TSR in descending order and place workload according to this ranked list. The CRAC are set to lowest TSR of all servers.

Dean *et al.* [4] first proposed Map-Reduce as a programming paradigm for distributed computing. They divide machines into one master node and several worker nodes. The master node chops data and passes the chunks to mapper nodes that generate intermediate results that are then aggregated on the reducers into the final result. Their original paper did not address energy concerns. Since then, much literature addressed energy efficiency in Map-Reduce. For example, Chen *et al.* [5], [6] discuss statistics of Map-Reduce workloads and suggest possible strategies to improve their energy behavior. Their experiments show that batching and staggered launching of batched jobs can help save energy. To the best of our knowledge, this paper is the first in considering the effect of the dependence of power consumption on temperature in the context of investigating energy-delay trade-offs of data-center workloads.

III. System Design

In this paper, we propose a general optimization framework for different applications, that helps to minimize service delay without violating a known power budget.

A. General Problem Description

In order to provide a general framework for optimization, we need to be aware of the controllable knobs in data centers. Since load balancers are often available in data-centers, the amount of workload w_i assigned to each machine i is one obvious knob. The workload w_i can be further divided into computational workload w_i^C and IO workload w_i^{IO} (*i.e.*, $w_i = w_i^C + w_i^{IO}$), and we assume that the ratio r between w_i^C and w_i is known for specific applications. Technologies like DVFS and Wake on Lan (WOL) help us enable two other widely used knobs: CPU frequency f_i and the number of powered-on machines n.

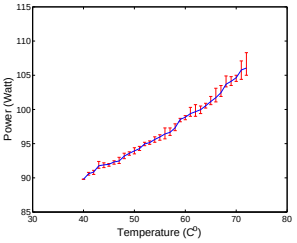


Fig. 1: Maximum Performance

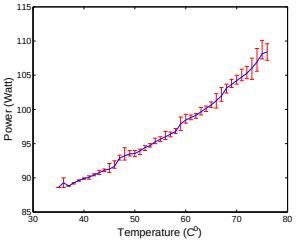


Fig. 2: Minimum Power

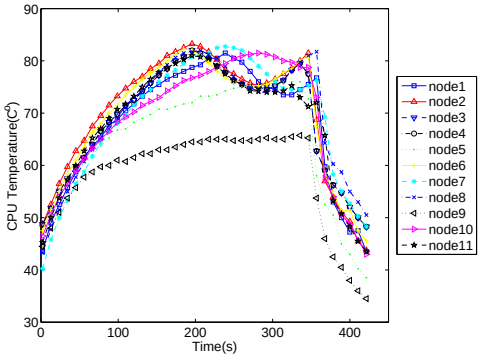


Fig. 3: Temperature Differences in Our Cluster: The cluster is running a computational intensive Map-Reduce job with 1 master node and 11 worker nodes. All machines locate in the same rack with different height. CPU frequencies of all worker nodes are set to the maximum value and the temperature set point of CRAC is set to 65 F^o .

Besides these three well studied knobs, we argue that CPU temperature T_i also exerts considerable influence on power consumption. Fig. 1 and Fig. 2 show the relationship between temperature and power in experiments where we run a CPU-intensive program to keep CPU utilization of the machine at 100%. The ambient temperature is then changed by controlling the set point of the CRAC unit. The CPU fan in our server (Dell PowerEdge R210) does not have a constant speed option. Instead, we measure the relationship between temperature and power in two cases; one where the fan policy was set to maximum performance (Fig. 1) and one where it is set to minimum power (Fig. 2). We can clearly see that, in both cases, power consumption increases as temperature rises, even though the load remains the same. In commercial data centers, heat is not uniformly distributed. The temperature differences can reach about 10 C^o in the same aisle [16]. The thermal gap can be even larger between CPUs in different server boxes. Figure 3 illustrates measured temperature variability in our tested. Hence, it is possible to improve energy savings by

better exploiting temperature differences. Thus, in this paper, we use (w_i, f_i, T_i, n) to formulate power consumption and delay.

The power consumption of one machine consists of two components: static and dynamic. The static component is incurred even when the machine is idle. The dynamic component depends on several parameters, such as CPU utilization, disk utilization, and CPU frequency. Our goal is to provide a general power model that applies to different applications. One server is a collection of several different components, such as, CPU, disk, memory, and NIC. The total power consumption of a server is the sum of components power draw. This nature of computer systems helps us to decouple the big problem into smaller ones, *i.e.*, we can model the power consumption of different components respectively and then, piece them together.

According to our experiments described in Section IV, the power draw of the CPU mainly depends on three detailed parameters: CPU utilization, CPU frequency, and CPU temperature. Equation 1 shows the relationship between the dynamic part of CPU power consumption and these three parameters when computationally-intensive workloads are running. Table I describes the meaning of related parameters used in this section.

$$P_{ij}^C = u_{ij}(a_1 f_{ij}^2 + a_2 f_{ij} + a_3)(T_{ij}^2 + a_4 T_{ij} + a_5) \quad (1)$$

CPU utilization represents busy CPU cycles over total CPU cycles. Given appropriate units, it can be represented as amount of computational workload over CPU frequency:

$$u_{ij} = \frac{rw_{ij}}{f_{ij}}. \quad (2)$$

In [33], Heath *et al.* proposed using a linear model of disk utilization to estimate disk power consumption. Inspired by this idea, we conduct experiments to measure the relationship between power consumption and disk Blocks-read-Per-Second (BPS). As we will discuss in section IV-B, our experiment result confirms validity of the linear model. Thus, we use equation (3) to model the dynamic part of disk power consumption.

$$P_{ij}^{IO} = a_6(1-r)w_{ij} \quad (3)$$

Since the two power models above contain only hardware status, they are not restricted to types of applications. Combining of equation (1) and (3) produces the general power model:

$$P_{ij} = P_{ij}^C + P_{ij}^{IO} + P^{idle}. \quad (4)$$

The purpose of our optimization is to calculate right states for controllable knobs such that the delay is minimized. Among the four parameters, CPU temperature is not directly tunable. Since the thermal state changes much slower comparing with computing. We propose to run optimizations periodically with the latest reported CPU temperature. Therefore, by substituting Equation (1) (2) (3) in 4, we have:

$$\begin{aligned} P_{ij} \quad &= w_{ij}(\alpha_{ij} f_{ij} + \beta_{ij} + \gamma_{ij}\frac{1}{f_{ij}}) + P^{idle} \\ where \quad &\alpha_{ij} = ra_1\widetilde{T_{ij}} \\ &\beta_{ij} = ra_2\widetilde{T_{ij}} + a_6(1-r) \\ &\gamma_{ij} = ra_3\widetilde{T_{ij}} \\ &\widetilde{T_{ij}} = T_{ij}^2 + a_4 T_{ij} + a_5 \end{aligned} \quad (5)$$

The power model applies to different kinds of applications. In contrast, the service delay highly depends on the nature of application themselves. For example, in a web service system where real-time request arrivals are Poisson and service time is exponentially distributed, the delay can be approximated by an M/M/1 model [32]. However, if we use Map-Reduce to process a large volume of data, jobs run in batch and the delay is simply modeled as the total amount of work over the resources the job occupies [34]. Given the diversity of applications, it is very hard to provide a general delay model. However, given a general delay function that is monotonically increasing with load and is a function of machines used and their frequency, different applications may share the same framework to determine optimal point. Since the knobs we choose are applicable among different cases, we let $D = G(\boldsymbol{N}, \boldsymbol{W}, \boldsymbol{F})$ denote the delay model where $\boldsymbol{N} = \{n_1, n_2, ..., n_l\}$, $\boldsymbol{W} = \{w_{ij} \mid i \le l, j \le n_i\}$ and $\boldsymbol{F} = \{f_{ij} \mid i \le l, j \le n_i\}$. Therefore, here we have the general optimization problem.

$$\begin{aligned} minimize \quad & D = G(\boldsymbol{W}, \boldsymbol{F}, \boldsymbol{N}) \\ s.t. \quad & \sum_{i=1}^{l}\sum_{j=1}^{n_i} P_{ij} \le P_b \end{aligned} \quad (6)$$

With the general formulation above, different applications may follow the same set of steps to calculate the global optimal:

1 Build a delay model, $G(\boldsymbol{W}, \boldsymbol{F}, \boldsymbol{N})$, according to application characteristics.
2 For one specific application, substitute static parameters with the numeric value to reduce unnecessary complexity. (*e.g.*, r equals 1 for workloads consuming only CPU resources, f_{ij} is a constant if the control policy does not tune CPU frequencies.)
3 If a convex problem is achieved in step 2, periodically conduct convex optimization to calculate and maintain optimal states.

In the following, we show the detail of how we apply our framework to two totally different applications and achieve global optimal.

B. Cloud Computing Workload

Hadoop is a widely used implementation of Map-Reduce in many cloud computing services. The performance objective of the target system in this case study is to minimize delay of serving one computationally-intensive job. According to [34], the delay is inversely proportional to the amount of resources

Symbol	Description
$\overline{D}$	average Service delay
a_*^*, b_*^*	Constant coefficient
W	Total workload
w_{ij}	Workload on jth server in tier i
r	Computational workload over total workload
n_i	Number of powered-on machines in tier i
D_{ij}	service delay on jth server in tier i
u_{ij}	CPU utilization of jth server in tier i
f_{ij}	CPU frequency of jth server in tier i
T_{ij}	CPU temperature of jth server in tier i
P_{ij}^{IO}	the Dynamic part of disk power consumption
P_{ij}^{C}	the Dynamic part of CPU power consumption
P_{ij}	Power consumption of jth server in tier i
P^{idle}	Power consumption when the server is idle
P_b	Power budget

TABLE I: Parameters Description

one job occupies. Hence, the overall delay of one job can be modeled as:

$$D \propto \frac{W}{\sum_{i=1}^{n_1} f_{1i}}, \tag{7}$$

Since there is only one tier of worker nodes, l equals to 1 in this case study. It is obvious to see that minimizing delay is the equivalent to maximizing the summation of CPU frequencies for all time units.

In Hadoop, the master node splits one job into many smaller tasks, and whenever one worker has any available resource slots, it will get one task to run with. This property means that the bottleneck resource on worker nodes will always be fully utilized. In our case, the CPU utilization is approximately 1. Since we are considering computational workload, the ratio r is also 1. Hence, according to Equation (2), the workload assigned to each machine can be represented as:

$$w_{1i} = f_{1i} \tag{8}$$

The purpose of our optimization is to find the right frequency for each machine such that the sum is maximized. The algorithm re-computes the solution periodically at fixed intervals. Note that, when a machine keeps running at high utilization, the CPU temperature does not change very quickly. Thus, each time when calculating the frequency for the current time period, we use the most recent reported temperature of each machine from the previous period as a constant to estimate the power consumption P_{1j}. In this way, we take the temperature dynamics into consideration in the equation, and, at the same time, we calculate the optimal frequency assignment taking temperature into account. Intuitively, we are using CPU frequency to compensate for the power contribution made by CPU temperature dynamics. Thus, substituting Equation (8) in (4), we get a new estimated power equation:

$$P_{1i} = \alpha_{1i} f_{1_i}^2 + \beta_{1i} f_{1i} + \gamma_{1i} + P^{idle}. \tag{9}$$

In this equation, the value of α_{1i}, β_{1i}, and γ_{1i} may vary for different machines due to their different individual CPU temperature as shown in 3. Thus, the optimization problem can be formulated as shown below.

$$\begin{aligned} maximize \quad & \sum_{i=1}^{n_1} f_{1i} \\ s.t. \quad & \sum_{i=1}^{n_1} P_{1i} < P_b \end{aligned} \tag{10}$$

The beauty of this problem is convexity. The objective function is linear. Since $\alpha_{1i} > 0$, each quadratic P_{1i} in the only constraint is positive definite and thus convex, so is their sum. Obviously, the definition domain is also convex. As a result, the problem is convex. Furthermore, the *Slater's* condition is also satisfied, because we can easily find at least one interior point in the feasible set after checking the constraint. Hence, the KKT condition is sufficient for the optimality of the primal problem. In other words, we can exactly obtain the optimum by solving the KKT conditions because the duality gap is zero.

Introducing Lagrange multiplier λ for the constraint, define:

$$L(\boldsymbol{F}, \mu) = \sum_{i=1}^{n_1} f_{1i} + \mu(\sum_{i=1}^{n_1} (\alpha_{1i} f_{1i}^2 + \beta_{1i} f_{1i} + \gamma_{1i}) - P_b). \tag{11}$$

Note that, the original problem described in Equation (10) has an optimal point at $f^*(t)$, if and only if there exists the optimal point ($\boldsymbol{F}^*$, μ) that maximizes Equation (11) with respect to $\boldsymbol{F}$. We then check the stationarity of the KKT conditions:

$$\frac{\partial L(\boldsymbol{F}, \mu)}{\partial \mu} = \sum_{i=1}^{n_1} (\alpha_{1i} f_{1i}^2 + \beta_{1i} f_{1i} + \gamma_{1i}) - P_b = 0,$$
$$\frac{\partial L(\boldsymbol{F}, \mu)}{\partial f_{1i}} = 1 + \mu(2\alpha_{1i} f_{1i} + \beta_{1i}) = 0.$$

Finally, from the 2 equations above, we get the optimal solution which also indicates the new frequency assignment for each machine.

$$\mu = \left[\frac{\sum_{i=1}^{n_1} \frac{1}{4\alpha_{1i}}}{P_b - \sum_{i=1}^{n_1} (\gamma_{1i} - \frac{\beta_{1i}^2}{4\alpha_{1i}})}\right]^{\frac{1}{2}} \tag{12}$$

$$f_{1i} = \frac{-1 - \mu\beta_{1i}}{2\alpha_{1i}\mu} \tag{13}$$

To calculate the optimal frequencies, we need to know the temperature of all machines in the cluster. Hence, there should be a global node that is responsible for the calculation. Fortunately, Map-Reduce itself provides the master node that collects information from all nodes through heartbeat messages. We add a temperature field in the heartbeat message and a frequency field in the return value. The size of heartbeat message may range from tens of bytes to hundreds of bytes

depending on different MapReduce cluster setup. Appending a four-byte float number will not lead to much network overhead. We discuss the implementation details in Section IV. The formulation above also indicates that the computational complexity is linear for both λ and f_{1i}. Thus, even though the optimization is done by a single centralized node, it will not consume too much computational resources. The performance evaluation is algorithm is presented later in Section V-A.

C. Enterprise Workload

The 3-tier web server farms are a widely used architecture serving many enterprise web e-commerce systems. The first tier offers the static contents in web pages and typically contributes the least to latency. The second-tier servers usually take care of web applications, while the third tier consists of a collection of database servers, which handle disk-intensive workload. In this section, we discuss an optimization that minimizes the delay of user queries while satisfying a given power budget. The incoming requests are uniformly distributed to all available second tier servers. The hosting application utilizes the Round-Robin balancer of JDBC to access third tier MySQL servers. Tuning both $\boldsymbol{F}$ and $\boldsymbol{N}$ makes the problem too complex to be solved by convex optimization. Therefore, we consider the more powerful knob $\boldsymbol{N}$. Let n_1 and n_2 denote the number of machines in 2nd-tier and 3rd-tier respectively. Therefore, we have:

$$\begin{aligned} w_{ij} &= \frac{W_i}{n_i} \\ \text{where} &\quad i \in \{1,2\}, j \leq n_i. \end{aligned} \tag{14}$$

According to our experiments shown in Fig.8 (b) and (c), M/M/1 model captures the service delay trends of both second and third tier servers. Thus, we use Equation (15) to estimate delay where b_{12} and b_{22} correspond to the service rate for second tier servers and third tier servers respectively.

$$\overline{D} = \sum_{i=1}^{2} \left(\frac{n_i b_{i1}}{n_i b_{i2} - W_i} + b_{i3}\right). \tag{15}$$

The definition domains of n_i is $(\frac{W_i}{b_{i2}}, \infty)$, which guarantees the system stability according to M/M/1 queueing model.

The second step is applying application properties to achieve convexity. Since we tune only the number of powered on servers, the CPU frequencies are constants in this case. The 2nd-tier servers handle computational workload that does not touch disk. Therefore we ignore power dynamics induced by disk for these servers. Further, by substituting Equation (14) in (4), we can reduce the power equation P_{1i}:

$$\begin{aligned} P_{1i} &= \theta_{1i}\frac{W_1}{n_1} + P^{idle}, \\ \text{where } \theta_{1i} &= \alpha_{1i} f_{1i} + \gamma_{1i}\frac{1}{f_{1i}} + r a_2 \widetilde{T_{ij}}, \end{aligned} \tag{16}$$

where θ_{1i} is the temperature dependent parameter. As we will present in Section IV-B, the CPU utilizations for 3rd-tier servers are always very low (less than 1 percent), we ignore the power dynamics induced by CPU for these servers. Further, by substituting Equation (14) in (4), we can reduce the power equation P_{2i}:

$$\begin{aligned} P_{2i} &= \theta_{2i}\frac{W_2}{n_2} + P^{idle}, \\ \text{where } \theta_2 &= a_6(1-r), \end{aligned} \tag{17}$$

For the purpose of optimization, we first regard n_i as continuous value, and then floor the optimal value to the nearest integer in implementation. Now, we formulate the optimization problem as follows.

$$\begin{aligned} minimize &\quad \overline{D} \\ s.t. &\quad \sum_{i=1}^{2}\sum_{j=1}^{n_i} P_{ij} < P_b \end{aligned} \tag{18}$$

We can evaluate that, in the definition domains of n_i, the Hessian matrixes of $\overline{D}$, and P_{ij} are positive definite, *i.e.*, the objective and constraint are all convex. Hence, the problem is convex. The Lagrangian, $L(\boldsymbol{N}, \mu)$, corresponding to our optimization problem can be expressed as follows:

$$\begin{aligned} L(\boldsymbol{N},\mu) &= \overline{D} + \mu[P_b - (\sum_{i=1}^{2}\theta_i W_i + P^{idle}\sum_{i=1}^{2}\theta_i)] \\ \text{where } \theta_1 &= \frac{\sum_{i=1}^{n_1}\theta_{1j}}{n_1} \end{aligned} \tag{19}$$

At the optimal point, the partial derivation of the Lagrangian with respect to n_i, and μ are zero. Thus, we get:

$$\begin{aligned} \frac{\partial L(\boldsymbol{N},\mu)}{\partial n_i} &= \frac{-b_{i1}W_i}{(n_i b_{i2} - W_i)^2} + \mu P^{idle} = 0, \\ \frac{\partial L(\boldsymbol{N},\mu)}{\partial \mu} &= P_b - (\sum_{i=1}^{2}\theta_i W_i + P^{idle}\sum_{i=1}^{2}\theta_i) \end{aligned} \tag{20}$$

From Equation (20), we get the optimal numbers of servers are:

$$\begin{aligned} n_i &= \frac{QW_i + K(b_{i1}W_i)^{\frac{1}{2}}}{Qb_{i2}}, \\ \mu &= \frac{Q^2}{K^2 P^{idle}}, \\ \text{where } K &= [\frac{P_b - \sum_{i=1}^{2}\theta_i W_i}{P^{idle}} - \sum_{i=1}^{2}\frac{W_i}{b_{i2}}]\prod b_{i2}, \\ Q &= b_{22}(b_{11}W_1)^{\frac{1}{2}} + b_{12}(b_{21}W_2)^{\frac{1}{2}}. \end{aligned} \tag{21}$$

Equation (21) presents the closed form solution for n_i. The time computation complexities are linear for two knobs, since we only have one summation equation to calculate θ_i.

IV. Implementation

In this section, we first specify the hardware used in our experiments. Then we demonstrate how we modify the Hadoop's heartbeat message. Finally, we show the experiments to derive power model and delay model respectively.

Fig. 4: Our cluster: there are 20 DELL PowerEdge R210 servers are available to use in this rack. Our experiments use 13 of them. When conducting experiments, all other machines are powered on and keep in idle status.

A. System Setup

Our evaluation involves 40 DELL PowerEdge R210 servers as shown in Fig. 4. Each server has a 4-core Intel Xeon X3430 2.4GHZ CPU. There are 11 available frequencies for each core: 1.197GHZ, 1.330GHZ, 1.463GHZ, 1.596GHZ, 1.729GHZ, 1.862GHZ, 1.995GHZ, 2.128GHZ, 2.261GHZ, 2.394GHZ, and 2.395GHZ. CPU turbo boost is turned off to avoid unpredictable dynamics. The server itself is equipped with CPU temperature sensors, and we collect the reading via lm_sensor module. There are two options available for the fan speed of our servers: minimum power and maximum performance. Both of them employed a closed feedback loop to dynamically determine proper fan speeds. As shown in Fig. 3, the characteristic of temperature power relationship does not change much for two fan configuration options. There is only a constant difference between them. Therefore, we choose the minimum power option in all our experiments to minimize the power consumed by fans. For the cloud computing case, we use $Hadoop-0.20.2$ as an implementation of Map-Reduce. One machine is nominated as a dedicated master node, *i.e.*, TaskTracker and other programs for executing tasks are not running on master node. For enterprise case, 18 dedicated 2nd-tier servers are installed with $Tomcat-6.0.32$ and 15 dedicated 3rd-tier servers are installed with $MySQL-5.5.15$. Besides, we have 3 dedicated Remote Browser Emulators (RBE), and one dedicated experiment controller which automates the experiment process and logs power consumption. We use one implementation of TPC-w benchmark designed by Wisconsin-Madison to populate our database [35]. Each MySQL server stores tables with 1×10^7 items (more than 5GB) by $MyISAM$ engine. The max_connection variable of MySQL is modified to tolerate up to 3000 connections. We use Avocent PM 3000 power distribution unit (PDU) to supply power for the rack, and directly collect readings from the PDU. The temperature of the server room is set to 65 F^o.

B. Power Equation

Because of temperature differences, the power consumption of one machine can be different even though it is running at the same frequency and same utilization. Understanding the relationship between CPU frequency, temperature and power consumption is crucial for conducting proper optimization. Fig. 5 shows how frequency and temperature influence power consumption.

Among, these three parameters, utilization and frequency can be well controlled by tuning workload and using DVFS. However, CPU temperature is not directly tunable. Actually, the CPU thermal states intricately relates to intake air temperature, air flow speed, location, CPU power draw, and etc, which is very hard to model. To walk around this, we tune the CRAC set point to affect CPU temperature. Then, we conduct 350 experiments that covers all combination of 10 frequencies (1.2GHZ $\sim$ 2.4GHZ), 6 utilization states (0 $\sim$ 100 percent), and 7 CRAC set points (50 $\sim$ 80 F^o). Finally, to plot the correlation of CPU power consumption with each individual parameter, we pick a subset of the data in which two parameters are roughly constant and the third one is changing. Fig. 5 shows the relationship between power consumption and three different parameters. The CPU utilization linearly relates the power consumption while both CPU frequency and CPU temperature show a quadratic relationship. To profile disk power consumption, we submit *SELECT* request to MySQL data base with different frequency. Fig. 8 (a) presents the linear relationship between disk block-read-per-second (br) relates and power consumption.

By piecing different parameters together, we get the following power model:

$$P = u \times (f^2 + 1.5253 \times f) \times (0.9937 \times T^2 - 1.3357 \times T + 3.2714) + 3.5321 \times br 40.5743,$$

As we argued before, for CPU intensive Hadoop workload load, the CPU utilization is always high. Thus, we pick the data set whose CPU utilization is higher than 99, and rerun the regression algorithm with a simpler target equation. Thus, we get:

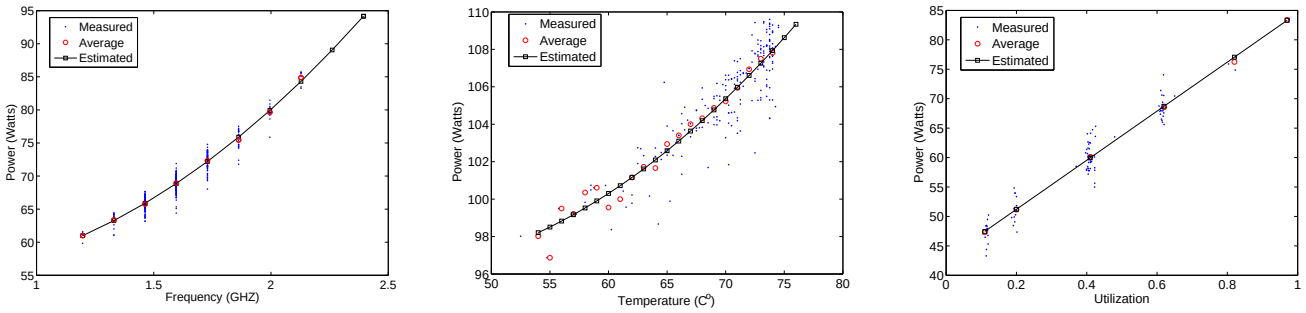
$$P = 0.0069 \times T^2 + 7.3865 \times f^2 + 0.0405 \times f \times T - 0.4351 \times T - 6.9793 \times f + 61.1205.$$

Regression result is shown in Fig. 6 which fits well with our observed data. The average error is about 0.106 watts which is relatively small comparing to the total power consumption of one machine.

In the enterprise experiment, by applying specific restrictions to power model, we have:

$$\begin{aligned} P_i^C = \;& 0.0782 \times w_i^C \times (0.9937 \times T_i^2 - 1.3357 \times T_i \\ & +3.2714) + 40.5743, \\ P_i^{IO} = \;& 0.021 w_i^{IO} + 40.5743. \end{aligned} \tag{22}$$

One thing to mention is that, when the MySQL server is handling disk-intensive requests, we observed that the CPU



(a) (b) (c)

Fig. 5: Correlation between power consumption and different parameters.

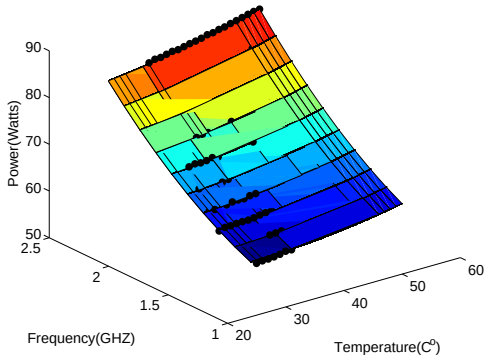


Fig. 6: Our power model and observed data: Every dot shows the average power for one temperature-frequency pair according to our measured data.

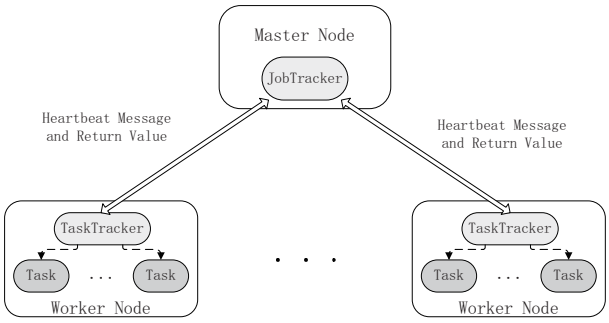


Fig. 7: Hadoop Cluster: There is one master node and several worker nodes in a Hadoop cluster. The JobTracker running on master node is responsible for task assignment while the TaskTracker running on worker node will launch tasks and report available slots to JobTracker. JobTracker and TaskTracker communicate with each other via Heartbeat message and its return value.

utilization is always less than 1 percent. Thus, we ignore the power consumption induced by CPU for third tier servers.

C. Modify Heartbeat Message

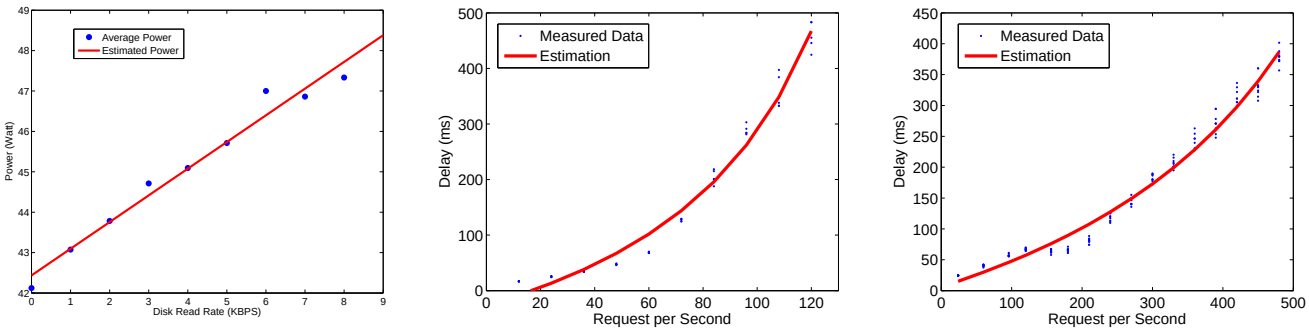
In the cloud computing case study, a Map-Reduce cluster consists of a master node and many worker nodes. More specifically, for Hadoop, there is a JobTracker program running on master node and a TaskTracker program running on each worker node as shown in Fig. 7. The TaskTracker sends heartbeat messages to inform JobTracker that the worker node is alive. In the heartbeat message, the TaskTracker will also indicate whether it is ready to run a new task. JobTracker responses a return value to TaskTracker which includes response ID, the interval for next heartbeat and probably information about a new task if the TaskTracker is ready. This structure grants global information to master node which is naturally suitable to run some centralized optimization algorithms.

We add a CPU temperature field to the heartbeat message. TaskTrackers sense the CPU temperature of its hosting machine and update this information before sending out heartbeat. In this way, JobTracker knows temperatures of all machines. JobTracker periodically calculates optimal frequencies for the cluster and sends it back via the return value of heartbeat. The calculation interval can be a tunable parameter which indicates the trade-off between optimization overhead and agility to dynamics. As discussed in section III-B, the computational complexity of our optimization algorithm is linear. Thus, it will not be a high overhead even if the cluster contains thousands of machines. The JobTracker only returns the optimal frequency and the frequency selection is done by worker nodes in order to reduce computation overhead on master node.

D. Delay Equation

In the enterprise workload case study, to capture the relationship between delay and load, we use one workload generator to produce *SELECT* queries, and record the average delay for every 100 request. Fig. 8 (b) and (c) show the experiment results of profiling delay-load relationship for second and third tier web servers. We use M/M/1 model to approximate the trends. Below show the detailed delay equation:

$$\begin{aligned} D_i^C &= \frac{16440}{62 - w_i^C} - 293, \\ D_i^{IO} &= \frac{10068}{300 - w_i^{IO}} - 330. \end{aligned} \tag{23}$$



(a) (b) (c)

Fig. 8: (a) Relationship betwen Power consumption and disk BPS for database workload. (b) (c) Relationship between delay and request-per-second for second and third tier servers respectively.

V. Evaluation

In this section, we evaluate the power consumption, service delay and reactions to dynamics for TAPA with respect to both MapReduce and web enterprise web servers.

A. TAPA and Cloud Computing Workload

Our evaluation includes five different settings, one with our temperature-aware DVFS algorithm enabled (TAPA), three static policies with CPU frequencies set to maximum, minimum and median respectively, and one temperature-oblivious DVFS policy proposed in [31]. All experiments involve 11 worker nodes, one master node, and one logging node which is responsible for log the real power readings from the power unit. We call them *TAPA*, *MAX*, *MIN*, *MED*, and *DVFS* below for short. For the DVFS one, we use their final control formulation shown in Equation (24) ,

$$f_i(k) = \frac{f_i(k-1)\sum_{T_{jl}\in S_i} c_{jl}r_j}{(U_s - u(k))f_i(k-1) + \sum_{T_{jl}\in S_i} c_{jl}r_j}. \quad (24)$$

Descriptions of all symbols are shown in Table II. Since this experiment focuses on computationally-intensive jobs, we set the U_s to 0.9. In Map-Reduce framework, master node will assign new task to a worker node whenever there is empty slot on that node, which keeps every core busy. Therefore, the estimated execution time for each processor can be set to the length of one control period (*i.e.*, $\sum_{T_{jl}\in S_i} c_{jl}r_j$ is replaced by a constant value). We apply a 1,000 watts power budgets to TAPA with 50 watts slack (*i.e.*, TAPA will try to keep the power consumption below 950 watts).

All of the experiments with different settings run the same computing PI job with 6,000,000,000 sample points, an example job provided by the original Hadoop-0.20.2 release. We configure each machine to hold 4 mapping slots since each of them has 4 cores. The mapper class here generates random points in a unit square and then counts points inside and outside the inscribed circle of the square. The reduce class accumulates point counts results from mappers and use this value to estimate PI. For the mapping phase, the cluster is computation intensive since all worker nodes are busy calculating and the traffic in the network is mostly heartbeat messages and its return value. When this job reaches its reducing phase, the network starts to become busy because many data need to pass from mappers to reducers.

Symbol	Description
$f_i(k)$	CPU frequency in the k^{th} control period
U_s	CPU utilization set point
$u(k)$	measured CPU utilization in period k
T_{jl}	subtask of Task T_j
S_i	set of subtasks located at processor P_i
c_{jl}	estimated execution time for subtask T_{jl}
r_j	calling rate of task T_j

TABLE II: Symbol descriptions for temperature oblivious DVFS

First, we show the temperature differences for all worker nodes of Med and TAPA experiment in Fig. 9 and Fig. 10 respectively. The settings of MED offers roughly the same computational capacity as TAPA since they take almost the same period of time to finish the same computational intensive job as shown in Fig. 14. Compared with Fig. 6, the thermal effect is even severe in this experiment. In MED experiment, the neighbors of each machine also generate a lot of heat. Therefore, the heat cannot be drained as easy as one machine cases. We can conclude at least two benefits from the comparison. Firstly, the maximum temperature difference between nodes in Med reaches 20 C^o, while it is only about 5 C^o in TAPA. It is obvious that TAPA achieves much smaller temperature deviation than Med. Secondly, the temperature of the hottest node reduces from 68 C^o to 48 C^o. Although we do not take removing hot spot as a direct objective, reducing energy cost by reducing temperature achieves this as a bonus effect. These two benefits are very meaningful in data centers. Smaller deviation means that the cooling part does not have to be over-provisioned. Operators can set the temperature target on CRAC to be just-enough for all machines. It can be a big saving from the cooling side. Lower temperature can help

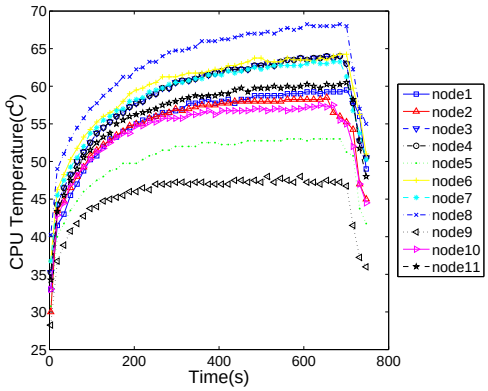


Fig. 9: Temperature differences for MED

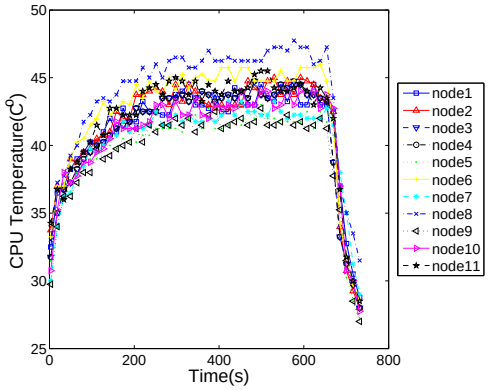


Fig. 10: Temperature differences for TAPA

to increase system stability and lengthen device life time. Therefore, it is another save in the long run. For each machine, there is a temperature drop in the tail. We think it is the time point when the map tasks have been finished and the whole cluster changes to the dedicated reduce phase. Therefore, at this time point, the job changes from computation intensive to communication intensive and the utilizations of CPUs decrease to low status.

In Fig. 11, we can see how our algorithm reacts to the temperature dynamics. This figure shows the relation between the frequency summation and the temperature trend. As shown in Fig. 10, different servers has similar temperature trend. Thus, we only use temperature curve from one machine to show this trend without making the figure too complicated. At the beginning, the temperature increases very fast because the cluster just switched from idle to full utilization status. TAPA immediately decreases the computational capacity of the whole cluster to curb power consumption below the budget. As the system keeps running, the temperature changes which will lead to different power consumption. In order to compensate temperature differences, our strategy decreases total CPU frequency when temperature increases and increases total CPU frequency when temperature decreases.

Finally, we show the comparison among TAPA, MAX, MIN,

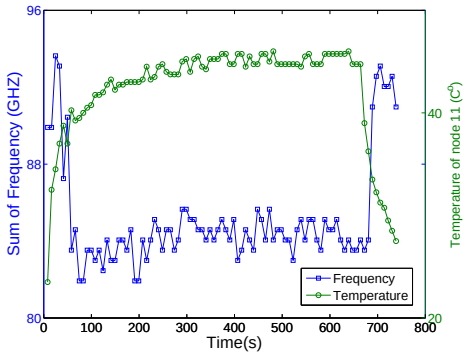


Fig. 11: Dynamic response: the curve with squares is the frequency summation of the whole cluster while the green curve shows the temperature dynamics of one machine which can represent the trends of the whole cluster.

MED, and DVFS based on their power consumption and efficiency. Fig. 12 shows the experiment result. The power budget is set to be 1000 watts with 50 watts slack for TAPA. For other approaches, the power budget does not apply to them since they do not have mechanism to control power consumption. From the Fig. 12, we can see that TAPA makes full use of the power budget without violating it. In MAX experiment, the power consumption rises as time passing by. It is because the machines gets hotter and hotter when running this computational intensive job. It is amazing that the power gap caused by temperature reaches about 200 Watts in our small cluster. The power consumption does not increase remarkably in MIN because machines are set to use the lowest frequency, with which the temperature can only reach about 34 C^o as shown in Fig. 6. For the MED experiments, the average temperature is about 55 C^o which leads to a 5 percent increase in total power. This also matches with our experiments result shown in Fig. 6. For the DVFS experiment, since all worker nodes stay in very high utilization status in map phase, the CPU frequency keeps increasing until it reaches the upper bound. Therefore, it shows a similar result as MAX. At the tail of each curve, there is a sudden drop of power. We believe it is because the cluster is transferring from mapping phase to reducing phase. And the CPUs no longer run at a high utilization.

Fig. 13 shows the frequency summation efficiency of five scenarios. The efficiency is defined as $\frac{F_s}{P}$, where F_s is the frequency summation of the whole cluster and P is the power consumption. It is clear that TAPA is more efficient to transform power into computational capacity when comparing with other solutions. As shown in Fig. 14, TAPA uses about 25% less energy comparing to the thermal-oblivious DVFS solution and about 5.6% less energy comparing to the most efficient static solution, MED. Given the huge expenditure spent on data center power supply, TAPA will help to achieve significant savings.

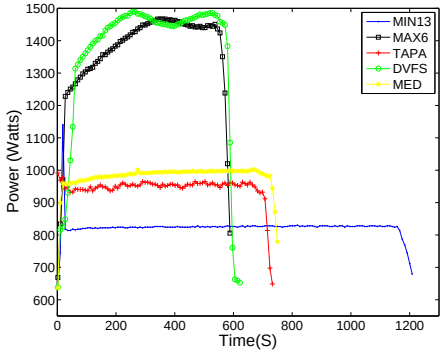


Fig. 12: Power consumption and running time

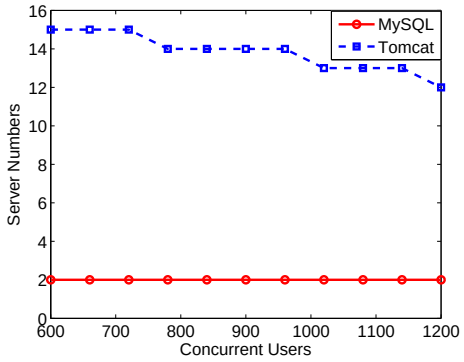


Fig. 15: Number of Servers

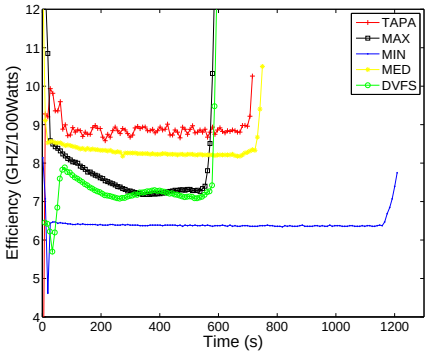


Fig. 13: Frequency summation efficiency

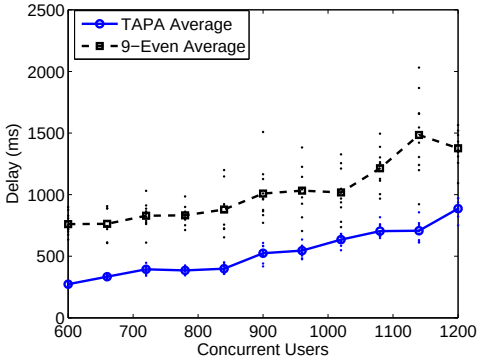


Fig. 16: Service Delay

B. TAPA and Enterprise Workload

In this section, we evaluate how our solution performs with web servers. The intuition behind TAPA is to allocation right number of servers for second and third tiers respectively, which minimizes service delay while satisfying a global power constraint. The experiment is designed in the way that W^C and W^{IO} are the same, and when cluster utilization is very low, delays induced by Tomcat server and MySQL server are roughly the same. But, as the volume of load goes up, the delay on Tomcat servers increases much faster than MySQL servers. We compare our solution with static allocation policy such that both second tier and third tier have 9 dedicated servers (9-Even). We set the RBE's average user think time to 1 second, and increase the number of users from 600 to 1200 with step 60.

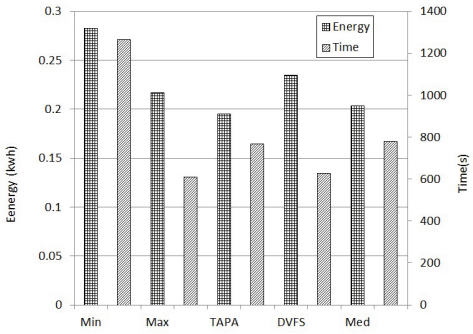


Fig. 14: Energy consumption and Job Duration

In Fig. 15, we show how TAPA reacts to as load increases. To curb the total power consumption, TAPA uses less machine to serve heavier load. Because the power induced by each machine rises with load, and we will have to reduce the amount of static power by turning off machines.

Fig. 16 and Fig. 17 together show the whole system power consumption and service delay for different numbers of users. We can clearly see that TAPA won in both aspects. In Fig. 16, even though 9-Even policy uses more servers, allocating even number of servers for each tier results in longer delay than TAPA. The reason is that, the bottleneck of the system is in 2nd-tier. Spending too much servers on 3rd-tier does not help to reduce delay. In Fig. 17, it is not surprising that 9-Even uses more power, since it is using more machines. One phenomenon we want to point out is that, the power consumption for 9-Even cases actually decreases as load goes up, we believe it is because when the arrival rate increases, the delay increases, which in turn reduces the number of request that the cluster can serve in each second. It does not change much the power draw of fully utilized Tomcat servers (Tomcat servers just spend more time on accepting request and switching between them). But, as the load goes to databases decreases, the power consumption of the 3rd-tier will also decrease.

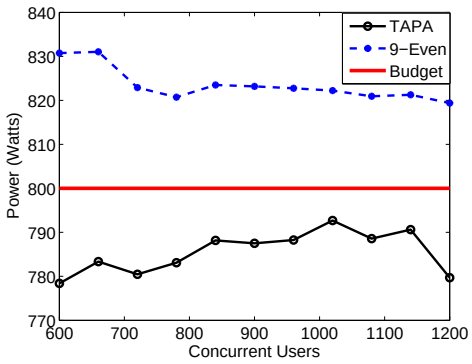


Fig. 17: Power Consumption

VI. Conclusion

In this paper, we present a general optimization framework for minimizing delay with a given power budget. Besides, we also provide a versatile power model that takes both CPU and disk into consideration. Based on the power model, we apply our optimization framework to two very different types of applications: Hadoop, and web servers. For each application, we only need to plug its specific delay constraints into the framework, and derive a closed form optimization formulation. The computational complexities for both cases are linear. The experiment results show that our solution can curb the power consumption very well in dynamic environments, which simultaneously reducing both power consumption and delay.

References

[1] James Hamilton, *Cost of Power in Large-Scale Data Centers*, http:// perspectives.mvdirona.com /2010 /09/ 18/ OverallDataCenterCosts. aspx, November 2008.

[2] Gartner, *Gartner Says Data Center Power, Cooling and Space Issues Are Set to Increase Rapidly as a Result of New High-Density Infrastructure Deployments*, http:// www.gartner.com /it/ page.jsp? id=1368614, May 2010.

[3] U.S. Department of Energy, *Quick Start Guide to Increase Data Center Energy Efficiency*. Retrieved 2010-06-10.

[4] J. Dean, and S. Ghemawat, *MapReduce: Simplified Data Processing on Large Clusters*, USENIX OSDI, December 2004.

[5] Y. Chen, T. Wang, J. M. Hellserstein, and R. H. Katz, *Energy Efficiency of Map Reduce*, http:// www.eecs .berkeley. edu/ Research/ Projects/ Data/ 105613. html, 2008.

[6] Y. Chen, A. Ganapathi, A. Fox, R. H. Katz, and D. Patterson, *Statistical Workloads for Energy Efficient MapReduce*, http:// www. eecs. berkeley. edu/ Pubs/ TechRpts/ 2010/ EECS- 2010-6. html, Jan, 2010.

[7] Hadoop. http://hadoop.apache.org, October 2011.

[8] Disco. http://discoproject.org, October 2011.

[9] Misco. http://www.cs.ucr.edu/∼jdou/misco, October 2011.

[10] P. Macken, M. Degrauwe, M. V. Paemel, and H. Oguey, *A voltage reduction technique for digital systems*. IEEE International Solid-State Circuits Conference, February 1990.

[11] W. Liao, F. Li and L. He, *Microarchitecture Level Power and Thermal Simulation Considering Temperature Dependent Leakage Model*, in Proceedings of International Symposium on Low Power Electronics and Design, pages 211-216, August 2003.

[12] L. Yuan, S. Leventhal, and G. Qu, *Temperature-aware leakage minimization technique for real-time systems*, ICCAD, Novoember 2006.

[13] L. He, W. Liao and M. R. Stan, *System level leakage reduction consider the interdependence of temperature and leakage*, in Design Automation Conference, 2004, pp. 12-17.

[14] W. Liao and L. He, *Coupled Power and Thermal Simulation and Its Application*, in the 3rd Workshop on Power-Aware Computer Systems, December 2003, pp.148-163.

[15] Y. Liu, R. P. Dick, L. Shang, and H. Yang, *Accurate Temperature-Dependent Integrated Circuit Leakage Power Estimation is Easy*, Conference on Design, automation and test in Europe, San Jose, CA, USA, 2007.

[16] C. M. Liang, J. Liu, L. Luo, A. Terzis, and F. Zhao *RACNet: A High-Fidelity Data Center Sensing Network*, ACM Sensys, November 2009.

[17] A. Gandhi, M. H. Balter, R. Das, and C. Lefurgy, *Optimal Power Allocation in Server Farms*, SIGMETRICS/Performance, Washington, USA, June 2009.

[18] C. Bash and G. Forman, *Cool Job Allocation: Measuring the Power Savings of Placing Jobs at Cooling-efficient Locations in the Data Center*, In Proc. USENIX Annual Technical Conference, Santa Clara, CA, June 2007

[19] T. Horvath, T. Abdelzaher, K. Skadron, and X. Liu, *Dynamic Voltage Scaling in Multi-tier Web Servers with End-to-end Delay Control*, IEEE Transactions on Computers, Vol. 56, No. 4, pp. 444-458, April 2007.

[20] V. Sharma, A. Thomas, T. Abdelzaher, K. Skadron, and Z. Lu, *Power-Aware QoS Management in Web Servers*, IEEE Real-Time System Symposium, December 2003.

[21] L. Wang and Y. Lu, *Efficient Power Management of Heterogeneous Soft Real-Time Clusters*, IEEE Real-Time Systems Symposium, December 2008.

[22] R. K. Sharma, C. L. Bash, C. D. Patel, R. J. Friedrich, and J. S. Chase. *Balance of Power: Dynamic Thermal Management for Internet Data Centers*, IEEE Internet Computing, 9(1):42-49, January 2005.

[23] J. Moore, J. Chase, P. Ranganathan, R. Sharma, *Making Scheduling "Cool": Temperature-Aware Workload Placement in Data Centers*, USENIX Annual Technical Conference, Berkeley, CA, USA, 2005.

[24] J. Leverich, and C. Kozyrakis, *On the Energy (In)efficiency of Hadoop Clusters*, HotPower, Big Sky, Montana, 2009.

[25] R. Kaushik, and M. Bhandarkar, *GreenHDFS: Towards An Energy-Conserving, Storage-Efficient, Hybrid Hadooop Compute Cluster*, HotPower, Vancouver, BC, Canada, October, 2010.

[26] W. Lang, and J. Patel, *Energy Management for MapReduce Clusters*, VLDB, Singapore, September, 2010.

[27] J. Heo, P. Jayachandran, I. Shin, D. Wang, and T. Abdelzaher, *OptiTuner: An Automatic Distributed Performance Optimization Service and a Server Farm Application*, International Workshop on Feedback Control Implementation and Design in Computing Systems and Networks (FeBID), San Francisco, California, April 2009.

[28] Z. Liu, M. Lin, and L. L. H. Andrew, *Greening Geographical Load Balancing*, SIGMETRICS, San Jose, California, USA, June 2011.

[29] A. Banerjee, T. Mukherjee, G. Varsamopoulos, and S. Gupta, *Cooling-Aware and Thermal-Aware Workload Placement for Green HPC Data Centers*, IGCC, Augest 2010, Chicago, IL, USA.

[30] M. Etinski, J. Corbalan, J. Labarta, and M. Valero, *Optimizing Job Performance Under a Given Power Constraint in HPC Centers*, IGCC, Augest 2010, Chicago, IL, USA

[31] X. Wang, X. Fu, X. Liu, Z. Gu, *Power-Aware CPU Utilization Control for Distributed Real-Time Systems*, Real-Time and Embedded Technology and Applications Symposium, San Francisco, CA, United States, 2009.

[32] M. Kihl, G. Cedersjö, A. Robertsson, B. Aspernäs, *Performance measurements and modeling of database servers*,Sixth International Workshop on Feedback Control Implementation and Design in Computing Systems and Networks (FeBID 2011), Karlsruhe, Germany, June 14, 2011.

[33] T. Heath, B. Diniz, E. V. Carrera, W. Meira, and R. Bianchini, *Energy Conservation in Heterogeneous Server Clusters*, In Proceedings of the Symposium on Principles and Practice of Parallel Programming (PPoPP), June 2005.

[34] K. Morton, A. Friesen, M. Balazinska, D. Grossman, *Estimating the Progress of MapReduce Pipelines*, IEEE International Conference on Data Engineering, Long Beach, CA, USA, March, 2010.

[35] http://pharm.ece.wisc.edu/tpcw.shtml

[36] Z. Liu, M. Lin, A. Wierman, S. Low, and L. Andrew, *Greening Geographical Load Balancing*, SIGMETRICS'11, NY, USA, 2011.