Reducing Data Center Energy Consumption via Coordinated Cooling and Load Management

Luca Parolini, Bruno Sinopoli, Bruce H. Krogh
Dept. of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh,PA 15213-3890
{lparolin|brunos|krogh}@ece.cmu.edu

ABSTRACT

This paper presents a unified approach to data center energy management based on a modeling framework that characterizes the influence of key decision variables on computational performance, thermal generation, and power consumption. Temperature dynamics are modeled by a network of interconnected components reflecting the spatial distribution of servers, computer room air conditioning (CRAC) units, and non-computational components in the data center. A second network models the distribution of the computational load among the servers. Server power states influence both networks. Formulating the control problem as a Markov decision process (MDP), the coordinated cooling and load management strategy minimizes the integrated weighted sum of power consumption and computational performance. Simulation results for a small example illustrate the potential for a coordinated control strategy to achieve better energy management than traditional schemes that control the computational and cooling subsystems separately. These results suggest several directions for further research.

Categories and Subject Descriptors

C.0 [**General**]: System architecture—*Data center, Energy management, Energy efficiency, Thermal modeling, Optimization, Optimal control, Cyber-physical systems*

1. INTRODUCTION

Data servers account for nearly 1.5% of total electricity consumption in the U.S. at a cost of approximately $4.5 billion per year [16]. Data center peak load exceeded 7 GW in 2006, and without intervention is destined to reach 12 GW by 2011, about the power generated by 25 baseload power plants. To address this emerging problem, this paper advocates a coordinated approach to data center energy management.

In most data centers today, the cooling and the computational subsystems are controlled independently. Computer room air conditioning (CRAC) units operate at levels necessary to keep the hottest regions in the data center below critical temperature limits. Computational tasks are allocated to achieve the desired performance with server power states that minimize overall server energy consumption. Some load management strategies consider the influence of the server loads on the temperature distribution, but CRAC control remains thermostatic. Coordinating cooling and load management requires a unified modeling framework that represents the interactions of these two subsystems. We present the elements of such a framework and formulate the coordinated control problem as a constrained Markov decision process (CMDP). Simulation results for a small system illustrate the potential benefits of this approach.

The following section reviews previous work on energy management strategies in data centers. Section 3 proposes an integrated thermal and computational model. Temperature dynamics are modeled by a network of interconnected components reflecting the spatial distribution of servers, CRAC units, and non-computational components in the data center. A second network models the distribution of the computational load among the servers. The server power states influence both networks. In Sec. 4, the energy management problem is formulated as a CMDP. Decision variables include the CRAC set points, the allocation of tasks to servers, and server power states. The dynamic MDP strategy determines these control variables as a function of the system state to minimize the integral of the weighted sum of power consumption and computational performance. Section 5 presents simulation results for a simple system, comparing the optimal MDP strategy to a conventional decoupled control of the cooling and computational subsystems. The final section identifies several directions for further research.

2. PREVIOUS WORK

Early work on energy minimization for data centers focused on computational fluid dynamic (CFD) models to analyze and design server racks and data center configurations to optimize the delivery of cold air, and thereby minimize cooling costs [13, 12]. Subsequent papers focused on the development of optimal load-balancing policies, at both the server and rack levels. Constraints on these policies were either the minimum allowed computational performance or the maximum allowed peak power consumption [5, 3, 4, 18, 6, 15].

Some recent papers have considered policies to minimize the total data center energy consumption by distributing the workload in a temperature-aware manner [11, 2]. Research on load distribution based on the thermal state of the data center, led to the development of fast and accurate estimation techniques to determine the temperature distribution in a data center, based on sparsely distributed temperature measurements [10, 14]. Approaches to the data center energy management problem based on queueing theory and Markov decision processes can be found in [9, 17].

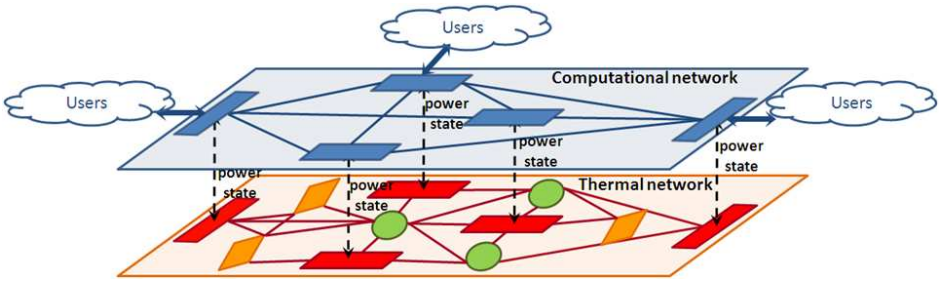


Figure 1: A layered data center model: a *computational network* of server nodes (rectangles); and a *thermal network* of server nodes (rectangles), CRAC nodes (circles), and environment nodes (diamonds). Server power states influence both networks.

3. INTEGRATED THERMAL - COMPUTATIONAL MODEL

As illustrated in Fig. 1, we model a data center as two coupled networks: a *computational network* and a *thermal network*. The computational network describes the relationship between the streams of data incoming from external users and generated by the data center itself and the computational load assigned to each server. The computational network is composed of *server* nodes that interact through the exchange of workloads. We call *jobs* the computational requests that enter and leave the data center and *tasks* the computational requests exchanged within the data center. A single job can be broken into multiple tasks and similarly, multiple tasks can be merged into a single job. A server's *workload* is then defined as the number of tasks arriving at the node per unit of time. The data center networks interact with the external world by exchanging jobs at the computational network level and through the consumption of electrical power at the thermal network level.

The thermal network describes the thermal energy exchange at the physical level of the data center and accounts for the power consumed by the CRAC units, IT devices and non-computational devices. The thermal network presents three types of nodes: *server*, *CRAC* and *environment* nodes. Environment nodes represent IT devices other than servers, non-computational devices and external environmental effects (e.g. weather). Each node in the thermal network has an input temperature $T\text{in}[^o\text{C}]$, an output temperature $T\text{out}[^o\text{C}]$ and an electrical power consumption $pw[\text{W}]$. The electrical power consumption is allowed to be either a controlled input or an output of a node (the result of some other controlled or uncontrolled processes).

Following the thermal model proposed in [14], we assume that the input temperature $T\text{in}_i[^o\text{C}]$ of a node i is given by a convex linear combination of the output temperatures of all other nodes and the value of each coefficient γ_{ij} depends upon the particular data center layout:

$$T\text{in}_i(t) = \sum_{j=1}^{N} \gamma_{i,j} T\text{out}_j(t) \ , \qquad (1)$$

$$\sum_{j=1}^{N} \gamma_{i,j} = 1 \ , \quad \gamma_{i,j} \geq 0, \qquad \forall\, i\,, j = 1, \ldots, N, \qquad (2)$$

where N is the total number of nodes in the thermal network. Details of the models for each type of node are given in the following subsections.

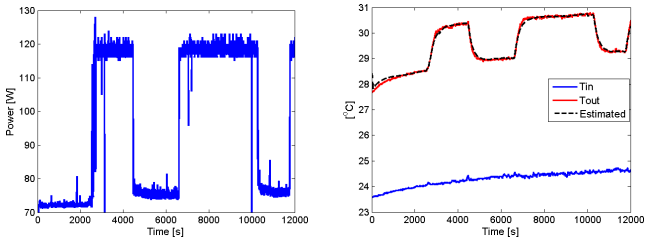


Figure 2: Empirical modeling of thermal server nodes. Left: Server node power consumption. Right: measured $T\text{in}, T\text{out}$ and predicted $T\text{out}$ from the empirical thermal server node model.

3.1 Server nodes

Each computational server node is associated to a unique thermal server node. We call server node the combination of them. The coupling between the two parts of a server node is given by the *power state* p. As proposed in [7], we assume that each server node has a finite number of possible power states denoted by the set P. A power state $p \in P$ determines uniquely both the mean execution rate $\mu[1/\text{s}]$ of each task and the power consumption $pw[\text{W}]$ of the node. We assume also that if $p_1, p_2 \in \mathcal{P}$ and $\mu(p_1) \geq \mu(p_2)$, then $pw(p_1) \geq pw(p_2)$.

We model the computational server node as a G/M/1 queue. The task inter-arrival distribution is given by a combination of the inter-arrival jobs at the data center and server interaction in the computational network, while the service time is exponentially distributed with parameter $\mu(p(t))$. Measurements taken on a server in our laboratory, shown in Fig. 2, show that the thermal part of server node can be modeled as a first-order linear time-invariant (LTI) system defined by the following differential equation:

$$\dot{T}\text{out}(t) = k(T\text{in}(t) - T\text{out}(t)) + c \cdot pw(t) \ . \qquad (3)$$

A server node can represent either a single server or a set of servers. In the second case it is necessary to identify aggregate definitions of the node power state, power consumption, input and output temperatures.

3.2 CRAC nodes

CRAC nodes reduce the input temperatures of other thermal nodes. Their inputs are: the input air temperature $T\text{in}$ and a reference output temperature value $T\text{ref}$; outputs are the output temperature $T\text{out}$ and the power consumption pw. When $T\text{ref}(t) \leq T\text{in}$, $T\text{out}(t)$ will tend to $T\text{ref}(t)$ according to the node dynamics, while $T\text{out}(t)$ will tend to $T\text{in}(t)$ when $T\text{ref}(t) > T\text{in}$. The electrical power consumption $pw(t)$ of CRAC nodes is assumed to be a function $f(\cdot,\cdot)$ of $T\text{in}$ and $T\text{out}$, monotonically decreasing in $T\text{out}$ for all $T\text{out} < T\text{in}$.

3.3 Environment nodes

Environment nodes model all other subsystems that influence the data center thermal dynamics, including IT devices that cannot be directly controlled, non-computational devices, and the external environment. The environment node output temperature is determined by a function $g(\cdot,\cdot)$ of $T\text{in}$ and pw. The temperature of the environment surrounding the data center can be expressed as the output temperature of an environment node taking $g(T\text{in}, pw) = T\text{in}$ and $pw = 0$[W].

3.4 Control Inputs

There are three sets of controllable variables in our proposed data center model: the computational network workload exchange, the server node power states and the CRAC node reference temperatures. Standard scheduling algorithms, in this framework, become controllers that try to optimize either the power consumption of server nodes or the power consumption of the whole data center, relying only on the information contained at the computational network level. Such an approach leads to sub-optimal solutions since the information given by the thermal network and consequently the possibility to act on a greater number of variables are discarded. Even thermal aware scheduling algorithms are sub-optimal since they do not take advantage of the possibility to control the CRAC reference temperatures.

4. CMDP FORMULATION

In order to synthesize data center energy management strategies, we use the model developed in the previous section to build a constrained Markov decision process (CMDP) [1]. The theory of CMDPs provides a powerful framework for tackling dynamic decision-making problems under uncertainty. Clearly our control problem can be naturally cast as a dynamic decision making problem under uncertainty. For example, we don't know when a task will arrive to a server node, or when it will be completed. CMDP techniques are the key to the synthesis of a controller when a system can be described by a finite set of states and a finite set of actions. In these cases the stochastic dynamic control problem can be solved using standard linear programming techniques.

For the sake of clarity, we show how to formulate the data center energy management problem as a finite CMDP using a particular instantiation of the data center model. Generalization to an arbitrary case is straightforward, but would require the introduction of more notation. Here we assume that the data center is composed of $n = 3$ server nodes, $r = 1$ CRAC nodes and $e = 0$ environment nodes. Also, we will consider a discrete-time model of the data center with time step $0 < \bar{\tau} < +\infty$.

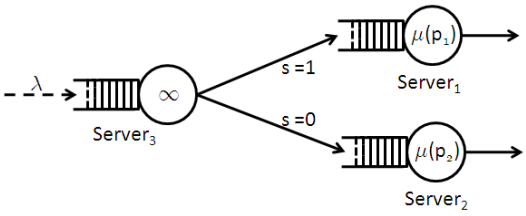


Figure 3: Example of computational network.

In order to formulate our optimization problem as a finite CMDP we have to identify: a finite set $\mathcal{X}$ of *states*, a finite set A of *actions* from which the controller can choose at each step $t = k \cdot \bar{\tau}$, a set $\mathcal{P}_{xay}$ of *transition probabilities* representing the probability of moving from a state x to a state y when the action a is applied, and a function $c : \mathcal{X} \times \mathsf{A} \rightarrow \mathbb{R}$ of *immediate costs* for each time step. The total cost over a given time horizon is the sum of the cost incurred at each time step.

For this example, we assume jobs arrive only through the third server node, called the *scheduler*, and there is no workload exchange between server nodes one and two, as illustrated in Fig. 3.

Jobs arrive at the scheduler one at a time at the beginning of every time slot k and the arrival events are i.i.d. Bernoulli random variables of parameter λ. The execution time of each job is independent from the execution time of all other jobs. During a single time slot a job is received by the scheduler and it is sent to one of the server nodes during the same time slot, according to a Bernoulli distributed random variable s. If $s = 1$ the job is sent to the server node 1, while if $s = 0$ the job is sent to the server node 2. The mean value of s can be chosen from a finite set of values.

In this example, there are no differences between jobs and tasks and the controller action on the computational network workload exchange reduces to the choice of the mean value of s.

The scheduler does not participate in thermal energy exchange between server nodes and CRAC node and its power consumption pw_3 is identically null for all of its power states. In the rest of the paper then, we will neglect its presence.

CRAC node is assumed to have no dynamics and hence, its output temperature coincides with $T\text{ref}$ when $T\text{ref} < T\text{in}$, while it is equal to $T\text{in}$ when $T\text{ref} > T\text{in}$.

In order to reach the goal of a finite set of states we have to discretize the set of admissible values for $T\text{out}_1$, $T\text{out}_2$ and $T\text{ref}$. We define $\tilde{T}\text{out}_1$ as the quantized version of $T\text{out}_1$ and similarly for $T\text{out}_2$. States of the CMDP will then be given by the pairs $(\tilde{T}\text{out}_1, \tilde{T}\text{out}_2)$, while actions will be given by tuple $(T\text{ref}, p \in \mathcal{P}_1, p \in \mathcal{P}_2, \mathsf{E}[s])$.

We model the uncertainty on $T\text{out}_i$ as a random variable uniformly distributed in $[\tilde{T}\text{out}_{i,min}(x), \tilde{T}\text{out}_{i,max}(x)]$. $\tilde{T}\text{out}_{i,min}(x)$ and $\tilde{T}\text{out}_{i,max}(x)$ represent respectively the minimum and the maximum temperature values allowed for

$T\text{out}_i$ when its discretized version has value $\tilde{T}\text{out}_i(x)$, while $\tilde{T}\text{out}_i(x)$ represents the value of $\tilde{T}\text{out}_i$ associated with the state $x \in \mathcal{X}$. The uncertainty about the true values of $T\text{out}_1$ and $T\text{out}_2$ determines each entry of $\mathcal{P}_{xay}$.

The selected cost function is:

$$c(x,a) = \sqrt{pw_1^2 + pw_2^2 + pw_4^2} + \alpha_\tau \tau_{max} \ , \qquad (4)$$

where pw_i represents the power consumption of the i-th node and τ_{max} is a function that depends on the mean waiting time of tasks. If the mean waiting time of a task is greater than a predefined threshold, then τ_{max} has value 1, otherwise it is set to 0.

4.1 Optimization constraints

The constraints of the optimization problem are the maximum and minimum input temperature allowed by the two server nodes: $T\text{in}_i \in [T\text{in}_{min}, T\text{in}_{max}]$, $i = 1, 2$.

Another constraint we add imposes that each server node has to be set to its most conservative power state if no tasks will be sent to it and that its mean task execution rate has to be greater than or equal to its task arrival rate.

4.2 Cost function

The cost function we want minimize is:

$$J(\mathbf{p}, \mathbf{T}\text{ref}, \mathbf{s}, \alpha_\tau, \tau_{max}) = (1-\alpha) \sum_{i=1}^{\infty} \alpha^i \mathsf{E}\Big[c(x(i), a(i))\Big], \quad (5)$$

where we denote with $x(i)$ and $a(i)$ respectively the state and the action at time i and with $\alpha \in (0,1)$ the *discount factor*.

This particular cost function can be called a *discounted energy*[1] cost since we are weighting the energy spent at the present time more than the one that will be spent in the far future.

5. SIMULATION RESULTS

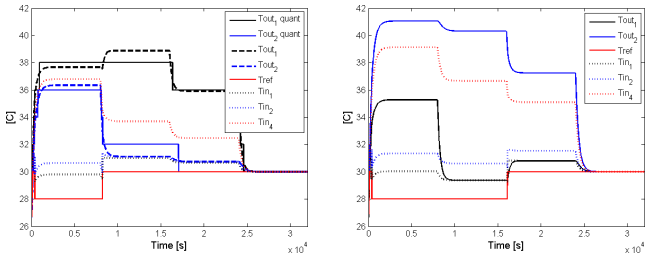
In order to solve the CMDP problem we use the Markov Decision Process Toolbox for MATLAB developed by the decision team of the Biometry and Artificial Intelligence Unit of INRA Toulouse, [8].

We compare our control algorithm against a typical hierarchical greedy scheduling algorithm that first chooses the best values for s, p_1 and p_2 to minimize the cost function given in equation (6) and then it selects the best value for $T\text{ref}$ in order to minimize the power consumption of the CRAC node enforcing the thermal constraints. The performances of the two algorithms are tested against four different job arrival rates: $\lambda = 0.9, 0.6, 0.2, 0$.

$$J(\mathbf{p}, \mathbf{s}, \alpha_\tau, \tau_{max}) = \sqrt{pw_1^2 + pw_2^2} + \alpha_\tau \tau_{max} \ . \qquad (6)$$

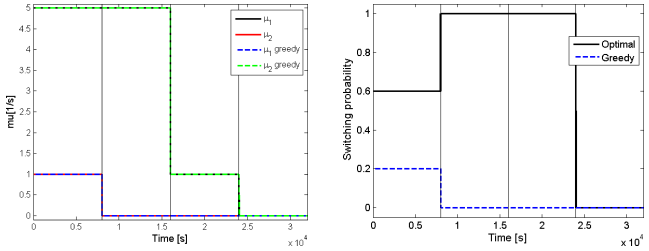
In this simulation we assume that server node 1 consumes more power than server node 2 for the same value of the task execution rate, while its thermal effect on the input temperature of the CRAC node is reduced with respect to the effects of server node 2. Finally we assume that the power consumption of the data center is mainly driven by the power of the CRAC node.

[1]When $\alpha_\tau = 0$ equation (5) represents the expected energy spent by the discretized system.



(a) Optimal controller case. (b) Greedy controller case.

Figure 4: Reference temperature, input and output temperature of each node.



(a) Task execution rate. (b) Switching probability.

Figure 5: Tasks execution rate and scheduler switching probability for both the optimal and greedy controller case. Black vertical lines are set correspondently to the instants where λ changes its value.

As shown in Fig. 4(a) and 4(b), using mainly server node one instead of the more power efficient server node two, the optimal controller is able to maintain a smaller difference between the input and output temperature of the CRAC node if compared to the greedy controller. Hence, as shown in Fig. 7(a), the optimal controller reaches a lower data center power consumption profile than the greedy controller and consequently a lower overall energy consumption, Fig. 7(b). As depicted in Fig. 6 energy minimization is obtained without heavily affecting the computational characteristics of the data center.

6. DISCUSSION

This paper presents a coordinated cooling and load management strategy to reduce data center energy consumption. This strategy is based on a holistic and modular approach to data center modeling that represents explicitly the coupling between the thermal and computational subsystems. This leads to a constrained Markov decision process (CMDP) formulation of the energy management problem for which optimal strategies can be computed using linear programming. Simulation results for a small example illustrate the potential for a coordinated control strategy to achieve better energy management than traditional schemes that control the computational and cooling subsystems separately. We are currently instrumenting a small data center at Carnegie Mel-

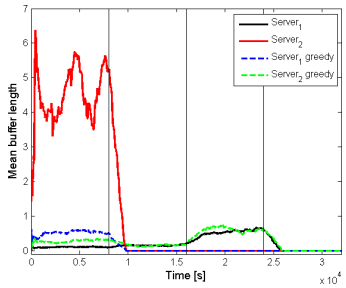
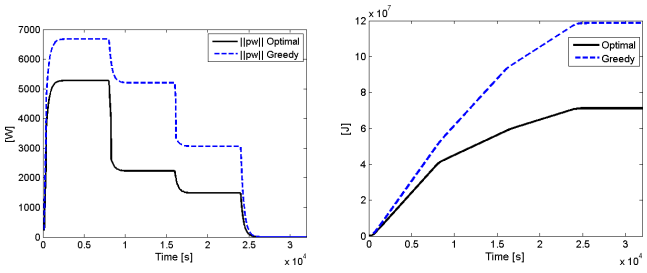


Figure 6: Mean buffers length. The mean is computed over a window of 30[min].



(a) Power consumption. (b) Energy consumption.

Figure 7: Comparison of energy and power consumption with optimal and greedy controller.

lon to test our algorithms in a realistic setting.

There are several directions for further research. The CMDP formulation will be intractable for detailed models of real systems due to the exponential growth of the state space. This problem can be addressed by using hierarchical models. In our modeling framework, the nodes in each network can represent aggregate sets of devices, each of which can be modeled by lower-level thermal and computational networks. We are currently developing the details of such hierarchical models. Hierarchical models will lead naturally to hierarchical and distributed control strategies. Initial steps in this direction can be found in [15].

Another issue regards the modeling and allocation of jobs and tasks in the computational network. Our current model assumes homogeneous and independent tasks and simple FIFO tasks queues at each server. In real systems, there can be a complex mixture of tasks and servers work on more than one task at a time. There may also be dynamic exchanges and reallocation of tasks that are in progress. We are currently investigating methods for incorporating and exploiting these features.

Finally, our model has several parameters that need to be determined for specific applications. This requires the development of effective system identification algorithms, which need to be implemented on line since the parameter values can be time varying.

7. REFERENCES

[1] E. Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, 1998.

[2] C. Bash, G. Forman. Cool job allocation: Measuring the power savings of placing jobs at cooling-efficient locations. Technical report, HPL, Aug. 2007.

[3] C.-H. Lien, Y.-W. Bai, M.-B. Lin, P.-A. Chen. The saving of energy in web server clusters by utilizing dynamic sever management. IEEE Int. Conf. on Networks, Nov. 2004.

[4] C. Lien, Y. Bai, M. Lin, C. Chang, M. Tsai. Web server power estimation, modeling and management. In IEEE, editor, *ICON '06. 14th IEEE Int. Conf. on Networks*, 2006.

[5] W. Felter, K. Rajamani, T. Keller, and C. Rusu. A performance-conserving approach for reducing peak power consumption in server systems. In *ICS*. ACM, 2005.

[6] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, F. Zhao. Energy-aware server provisioning and load dispatching for connection-intensive internet services. In *NSDI'08*.

[7] Hewlett-Packard Corporation, Intel Corporation, Microsoft Corporation, Phoenix Technologies Ltd.,Toshiba Corporation. Advanced configuration and power interface specification. Technical report, 2005.

[8] I. Chadès, M. Cros, F. Garcia, R. Sabbadin. Markov Decision Process (MDP) Toolbox v2.0 for MATLAB. www.inra.fr.

[9] L. Mastroleon, N. Bambos, C. Kozyrakis, D. Economou. Autonomic power management schemes for internet servers and data centers. In *IEEE GLOBECOM*, Nov. 2005.

[10] J. Moore, J. Chase, and P. Ranganathan. Consil: Low-cost thermal mapping of data centers. June 2006.

[11] J. Moore, J. Chase, P. Ranganathan, and R. Sharma. Making scheduling "cool": temperature-aware workload placement in data centers. In *ATEC*, 2005.

[12] Patel C. D., Bash C. E., Sharma R. K, Beitelmal A, Friedrich R. J. Smart cooling of datacenters. July 2003.

[13] Patel, C.D., Sharma, R.K, Bash, C.E., Beitelmal. Thermal considerations in cooling large scale high compute density data centers. May 2002.

[14] Q. Tang, T. Mukherjee, S. K. S. Gupta, P. Cayton. Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters. Oct. 2006.

[15] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, X. Zhu. No "power" struggles: coordinated multi-level power management for the data center. *ASPLOS*, 2008.

[16] U.S. Environmental Protection Agency. Report to congress on server and data center energy efficiency. Technical report, ENERGY STAR Program, Aug. 2007.

[17] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, N. Gautam. Managing server energy and operational costs in hosting centers. *SIGMETRICS*, 2005.

[18] Z. Xue, X. Dong, S. Ma, S. Fan, Y. Mei. An energy-efficient management mechanism for large-scale server clusters. In *APSCC '07*, 2007.