Dynamic hosting management of Web based applications over clouds[1]

Zahra Abbasi*, Tridib Mukherjee†, Georgios Varsamopoulos* and Sandeep K. S. Gupta*
*Impact Lab, School of Computing, Informatics and Decision Systems Engineering, ASU, Tempe, AZ
{zahra.abbasi,georgios.varsamopoulos,sandeep.gupta}@asu.edu
†Xerox Research Center, Webster, NY
Tridib.Mukherjee@xerox.com

***Abstract*—Dynamic Application Hosting Management (DAHM) allows clouds to dynamically host applications in data centers at different locations based on: (i) spatio-temporal variation of energy price, (ii) data center computing and cooling energy efficiency, (iii) Virtual Machine (VM) migration cost for the applications, and (iv) any SLA violations due to migration overhead or network delay. DAHM is complementary to dynamic workload distribution problem and is modeled as *mixed integer programming*; online algorithms are developed to solve the problem. The algorithms are evaluated in a simulation study using realistic data and compared with performance-oriented application assignment, i.e., hosting the application at a data center whose delay is the least. Our simulations results indicate that DAHM can potentially save up to 20% cost while incurring only a nominal increase in SLA violations. The savings are obtained by exploiting the cost efficiency variation as well as reducing the total number of VMs employed to host applications.**

***Keywords*-cloud computing, application hosting, Data center power efficiency**

I. Introduction

With the increasing prevalence of Internet-based computing services such as online gaming [1], cloud based services [2], and search engines, the energy consumption in data centers to host such services has skyrocketed and so has the importance of saving energy in data centers. According to a report by Intel and Microsoft [3], the energy cost accounts for over 10% of the total cost of ownership (TCO) of a data center.

Cloud computing is an emerging paradigm based on virtualization [4] which facilitates a dynamic, demand-driven allocation of computation load across data centers. Applications can be assigned to *Virtual Machines (VMs)* independent of physical infrastructures. Virtualization provides a cloud flexibility to serve an application on the most cost efficient data center at the time.

The *cost efficiency of data centers* may vary depending on several factors [5], [6]. First, the electricity price may vary over location and time (see Fig. 2). Second, *Power Usage Effectiveness (PUE)*, which measures the efficiency of a cooling system and any source of power consumption other than the computing equipment in data centers, may vary for different data centers [7], [8]. For example, according to [9], on average, data centers have a PUE of 1.7, whereas Google's modern data centers have a PUE of 1.18 [10]. Finally, the computing power consumption model of data centers (power consumption over utilization of the servers) may vary depending on the power behavior of the servers. Servers should be *ideally energy-proportional* [11], [12], [13], [14], i.e., they should have zero power at idle and a linear increase of power with respect to utilization. However, in reality, servers do not have zero power at idle [14]. Much work has been performed to develop power state transitioning and server provisioning schemes to remove idle power [15], [16].

In this paper, we propose *Dynamic Application Hosting Management (DAHM)* that uses the flexibility of cloud computing to decrease the cost of hosted applications by considering: *(i) the power consumption behavior and the energy cost of data centers within a cloud*, and (ii) *applications' performance requirement and bandwidth cost of live migration.* The spatial and temporal variation of electricity cost and data centers' average performance have been previously considered in developing workload placement algorithms [6], [17], [18]. However, our model improves the cost saving of previous approaches when using non energy-proportional servers. Under workload distribution management scheme that does not take into account energy proportionality of servers, only the utilization energy can be reduced, whereas VM management reduces total power including the VMs' overhead. Table I lists how the cost parameters can be considered in VM and workload management and whether they affect the total cost and the SLA violations.

In addition to the aforementioned cost efficiency factors, dynamically hosting applications in data centers also requires awareness of the network delay and the bandwidth overhead during VM migration. This overhead depends on the type of Web applications, which can be either stateless or stateful. In *stateless applications*, e.g., search engines, the state of online users are not recorded; whereas *stateful applications*, e.g., multi-player online games and cloud based services, keep track of the state of users [1]. To manage the hosting of stateless applications, their persistent data are

[1] This work was supported in part by grants from Intel Corp., NSF CRI grant #0855527, and NSF grant CNS/II-EN-0855277.

[†] The work was done when the author was at Arizona State University.

Table I: VM management versus workload management.

This cost parameter	can be considered in	and affects
utilization power	workload and VM mgt.	total cost
Idle power	VM mgt.	total cost
electricity price	workload and VM mgt.	total cost
workload variation	workload and VM mgt.	total cost
migration overhead	VM mgt.	total cost, SLA
network and service delay	workload and VM mgt.	total cost, SLA

replicated over the cloud, and multiple instances of the application can be activated across the cloud without the need for data migration. In case of stateful applications, if their instances migrate, their corresponding state information needs to be moved as well. Consequently, stateful applications tend to higher VM migration cost. All related work approaches are designed for stateless applications, where migration costs are negligible.

For Web applications, requests may originate from different locations. As such, the network delay from these locations to the hosting data centers might also impact the delay of the application. This may prevent an application from being hosted at certain locations. The population of online users for a specific Web application varies over time and has a cyclic behavior (i.e., hourly, daily, weekly. etc.) with respect to their local time. For instance, the Fig. 3, demonstrating the number of online users for an entertainment application, shows that the peak traffic time for different locations does not occur at the same time.

The work in this paper takes into account the cost efficiency factors, the network delay, the migration overhead and the traffic behavior in the formulation of the DAHM problem, and provides offline optimal solution and online heuristics to determine in which data centers to host the application's VMs. To this end, the following contributions have been made:

- *Modeling of DAHM problem*: DAHM is formulated as a *mixed integer programming (MIP)* formulation. The objective of the problem is a cost model for DAHM that accounts for energy cost of data centers, cost due to performance violation, and migration cost.
- *Online DAHM solutions*: We provide optimal offline solution by using the GLPK solver [19] for the MIP problem and develop greedy online heuristics that dynamically host application's VMs based on the cost of data centers (the computation complexity of the greedy algorithm is $O(|S||A|)$, where $|S|$ and $|A|$ are the total number of data centers and the total number of areas, respectively).
- *Simulation based evaluation*: The cost efficiency of the algorithms is shown through a simulation study under realistic workloads, electricity pricing and data center power performance. Results show that using DAHM online solutions, the cost can be saved up to 20% in the tested cases, depending on the data center power consumption model and heterogeneity of data centers. The cost saving under non power-proportional server interestingly increases due to the reduction in the number of VMs across data centers.

The rest of the paper is organized as follows. Section II presents the related work. Section III presents the system model including cost model, performance model, and DAHM problem formulation followed by the online algorithms for DAHM in Section IV. Section V presents the simulation based evaluation of DAHM and finally, Section VI concludes the paper.

II. Related Work

The cloud's virtualization technology and the spatio-temporal variation of electricity price have been leveraged in recent times to design cost efficient workload placement algorithms [6], [18], [17], [20], [21]. Rao et al. [6] consider the load distribution across data centers with the objective of minimizing the current energy cost subject to delay constraints. They use linear programming techniques and min-cost flow model to solve the problem and show the cost efficiency of their approach through simulation. The authors further improve their scheme by proposing a joint optimization of power control and electricity price in [20].

Another scheme for workload scheduling across data centers has been developed in [18], where the problem is modeled as a non linear optimization problem and solved using simulated annealing. Their results show that by leveraging the electricity price, significant cost can be saved when servers are assumed to be ideally energy-proportional, and cost saving decreases when servers have idle power greater than zero.

Qureshi et al. [17] used heuristics to quantify the potential economic gain of considering electricity price at the location of computation. Through simulation using realistic historical electricity price and workload traces, they reported that judicious location of computation load may save millions of dollars in the total operation costs of data centers. They also showed that the amount of cost saving depends on how energy-proportional the servers are and whether there is any constraint on the network bandwidth. They found that the cost saving is the highest when servers are ideally-energy proportional and all the network bandwidth is unconstrainedly available.

The DAHM solution can complement the aforementioned research because of the following reasons: first, the migration cost needs to be incorporated into the optimization problem, as it can be a significant factor for stateful applications; second, all related work show the significant impact of servers' idle power in the cost saving of workload distribution across data centers. Nevertheless, the energy proportionality of a server depends not only on the idle power, but also on the

Table II: Symbols and definitions.

Symbol	Definition
t	Epoch index
i	index of data centers
j	index of areas
$x_{i,j,t}$	the assignment of area j to data center i at time t
τ	the length of epochs (in second)
p_i^{idle}	The idle power of a server at data center i
p^{util}	The peak power minus p^{idle} at data center i
c_i	the avg. utilization of a user on a server at data center i
u	utilization
u_i^{th}	threshold utilization of servers at data center i
n_i^{VM}	number of active VMS at data center i
s_i	number of available VM at data center i
d	total delay of of a request
d'	data center delay
d''	network delay between areas and data centers
d^{ref}	total reference delay
d'^{ref}	service delay
e_i	electricity cost at data center i
β	Migration cost per each migration
α	Switching cost per turning on of a new server
η	performance violation cost per each user

power management schemes of data centers [15], [16], i.e., server provisioning and server-level power state transitioning. For that, we provide modeling and results to consider the effect of power management algorithms, i.e., server provisioning across data centers. Finally, the delay sensitivity of applications that may prevent them from being run at certain locations should be taken into account. Previous work modeled the data centers' response time from the service front-end's point of view [18], [6], thus the network latency between the clients and data centers are missing. Qureshi et al. [17] showed that it is important to consider the bandwidth cost of the network cost between the end users and data centers, however they did not provide modeling and optimal solutions of the problem.

III. System Model

In this section, we present the system model and formally define the DAHM problem.

We assume that a Web application can be hosted on a cloud consisting of several data centers. Let $S_t = \{s_{1,t}, s_{2,t}, \dots\}$ be the set of data centers available in the cloud, each $s_{i,t}$ represents number of available virtual machines (VMs) in data center i at time t. An application can be hosted in one or more data centers at the same time. In other words, depending on the application, replication can be allowed. Also, within a data center, the application can be either assigned to one or more VMs. We define a cost function that can model energy cost of all of these options. Also, another basic assumption of our model is that electricity price may vary over time and location. Therefore, data centers' energy cost varies with respect to their physical location and time of the day.

A. Performance modeling

We assume the quality of service of Internet applications to depend on *delay*. For end users to experience a high quality of service, their delay should not go above a *reference delay*, d^{ref}, whose value depends on the application. The delay experienced by a user consists of service delay, i.e., data center delay, d', and network delay, i.e., the delay between the user and the data center, d'' such that the total delay becomes: $d = d' + d''$.

In Internet data centers, the SLA statistically bounds the delay, i.e., the delay of q percent of requests should not go beyond the reference service delay d'^{ref}. We borrowed the performance model provided by [22], [23] to guarantee the SLA for data centers. This model asserts that the servers' utilization level is strongly correlated to the service delay; specifically, the SLA is violated when the utilization goes above a reference point [22], [23]. The threshold point depends on the hardware capacity of servers and the type of requests. We consider that there is an associated u_i^{th} for each data centers' servers, such that if servers are not utilized above that point the service delay (i.e., d_i') respect the SLA (i.e., $d_i' \leq d^{ref}$ for q percentage of users).

To model the network delay, we consider that the network delay depends upon the network between the end users at different areas and the data centers. Therefore, to consider the delay constraint, we model the geographical location of end users. Let A be the set of non-overlapping areas i.e., $A = \{a_1, a_2, \dots\}$. Users in area a_j receive service from the data center s_i with delay of $d''_{i,j,t}$ at time t. This delay can vary over time depending on the network congestion. Hence, to guarantee the SLA, $d_{i.j.t} = d'_{i,t} + d''_{i,j,t} \leq d^{ref}$ for q percentage of users.

B. Workload modeling

The set of online users at a particular area can represent a job that can be assigned to an instance of application running in a data center. Hence, we assume that users of an area are assigned to only one data center. Let N_t be the set of number of online end users in each area where $N_t = \{n_{1,t}, n_{2,t} \dots n_{|A|,t}\}$. The set N_t varies over time because first, different applications have different local peak times during a day; second, due to the time zone difference across geographical areas, the peak time of an application is different in these areas. The physical hosting location of an application may also vary over time, which may need live migration of the application. If a data center hosts a VM of the application to serve a certain area, then the *area is said to be assigned to the data center*. Live migration imposes some costs including the network usage, delay and the possible setting up of a new instance of the application in data centers which should be taken into account in the DAHM problem.

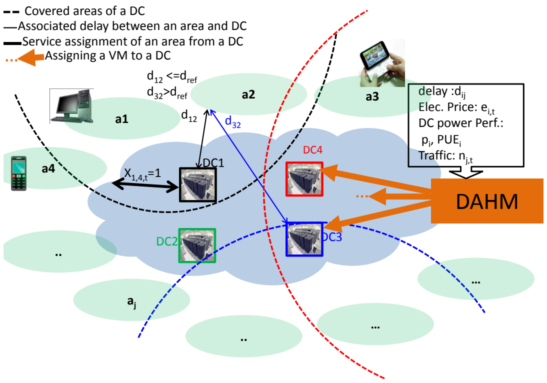

Figure 1: Pictorial view of system model.

We assume that the distribution of users and of electricity price vary over time and space. However, these remain constant within a time epoch (say half an hour, or one hour). To formally define the decisions of the manager, we can change the assignment of online users in an area j to the instance of application hosted in data center i at time t by an indicator variable $x_{i,j,t}$, then the goal is to determine the vector $\mathbf{x}$ in a discrete time system, where each time step $t \in \{1, \ldots, T\}$ represents a different *epoch* of length τ.

C. Energy cost

We assume that the energy cost of hosting an application in a data center is a function of the power performance of the VMs inside a data center, cooling energy and the electricity price of the data center.

The power model of a server in a data center specifies how the server consumes power with respect to its utilization (i.e., average utilization for all components). Servers in a data center can be non energy-proportional and heterogeneous (i.e., the power consumption of servers can be different). Finally, the power management schemes of data centers affects the server's power consumption [15], [16].

To capture these properties, we model the power consumption of a server in a data center which is allowed to be non energy-proportional as follows: $p_{i,t} = u_{i,t}p_i^{util} + p_i^{idle}$, where p_i^{idle} is the average idle power consumption of a server in data center i, p_i^{util} is the extra power consumption of a server at full utilization, and $u_{i,t}$ represents the utilization of servers of data center i from running the application at epoch t.

The utilization of a server depends on its workload and its physical characteristics. Workload of a server hosting an application is a function of its online users. Therefore, we assume the following linear model for utilization of a system: $u_{i,t} = c_i n_t$, where $u_{i,t}$ denotes the average utilization that users of area j impose to the data center i during epoch t, c_i is the average utilization that one online user imposes on the data center i during one second which depends on the system hardware characteristics and type of the application, and finally $n_t = \sum_{j=1}^{|A|} x_{i,j,t}n_{j,t}$ is the total number of users that are assigned to the data center i at epoch t. This model is frequently used in existing literature and experimental results show its sufficiency [22].

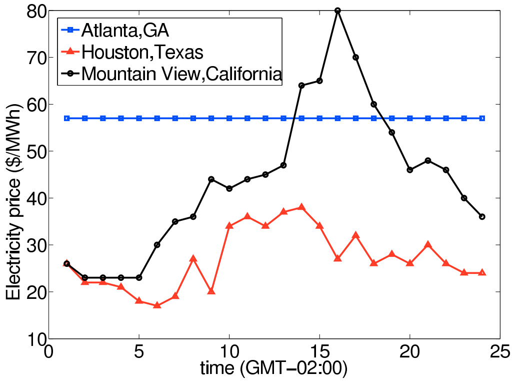

Figure 2: Hourly electricity price data for three major locations of Google IDCs on May 2nd, 2009 (data are taken from [6]).

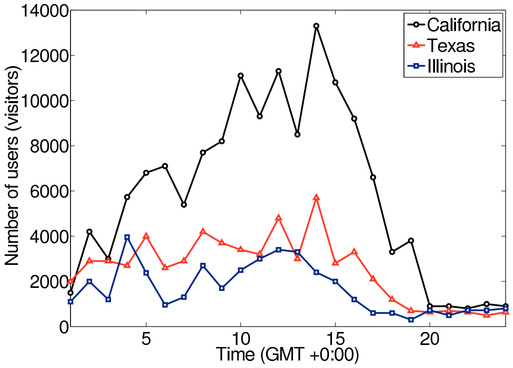

Figure 3: Hourly number of online users from three states for an entertainment Web site hosted at Go Daddy.

The power consumption of a server running a VM can be modeled with the same model of a physical server. In other words, the power consumption of a VM can be non energy-proportional as it consumes a significant of power when the VM is idle [24]. When number of online users increases, more than one VMs may be allocated to the application. Assume n_i^{th} is total number of users that a single VM in data center i can afford such that $c_i n_i^{th} = u_i^{th}$, then the idle power consumption equals to $n_{i,t}^{VM} p_i^{idle}$, where $n_{i,t}^{VM}$ is the ceiling integer part of n_t/n^{th}.

Power management schemes can be implemented at the server or the data center level: (i) server level power state transitioning [15], i.e., transitioning to the sleep state when no workload is available, and waking up on receiving a workload, and (ii) server provisioning [16], i.e., adjusting the number of active servers to the offered workload. Server level power state transitioning makes servers more energy-proportional; for that, we provide numerical results in the simulation section to show its effects. Server provisioning also incurs a power cost for switching on a server or launching a new instance of VM [24], [25]; for that, we consider that the constant cost of α_i is incurred whenever a new instance of VM is launched at data center i.

Non computing equipment power consumption, e.g., cooling power, can be estimated as a product of PUE and the total computing power of a data center.

To calculate the energy cost during an epoch, it suffices to multiply the power consumption into electricity price at the data center i of time t. We denote the electricity price by $e_{i,t}$, so the total electricity price of an application hosted during an epoch t in a data center i equals to:

$$Cost_{i,t}^{energy} = \left(\sum_{j=1}^{|A|} x_{i,j,t} n_{j,t} c_i p_{i,t}^{util} + n_{i,t}^{VM} p_i^{idle} \right) PUE_i e_{i,t} \tau \\ + \textstyle\sum_{i=1}^{|S|} \alpha_i (n_{i,t}^{VM} - n_{i,t-1}^{VM})^+ \tag{1}$$

Note that the '+' in the above equation points that the cost is non-zero *only* when the difference is positive. According to the above model, when no online users are assigned to a data center i, all the application's associated VMs (if there are any) are deactivated; consequently the idle power cost as well as the utilization power cost become zero.

D. Performance cost

To take into account the delay requirement of applications we make a performance violation cost model as $\text{cost}_{i,t}^{\text{perf. viol.}} = \sum_{j=1}^{|A|} \eta n_{,t} x_{i,j,t} (d_{i,j,t} - d^{ref})^+$. According to this model, the DAHM considers a punishment value of η for each user whose delay requirement is not met.

E. Migration cost

The DAHM may need to migrate an application (i.e., its state information) from a data center to another. However, migration imposes cost in terms of increase in network bandwidth consumption, delay on current online users, and power consumption for the possible setting up of a new instance of the application. For simplicity, we assume all of these costs can be aggregatedly denoted by a constant β, incurred whenever a migration happens. From the model, it follows that migration happens if users who were previously not assigned to a data center i (i.e., $x_{i,j,t-1} = 0$), in the current time are assigned to it (i.e., $x_{i,j,t} = 1$). Therefore, the migration cost for a data center i at time t can be formulated as $\text{cost}_{i,t}^{\text{migration}} = \beta \sum_{j=1}^{|A|} (x_{i,j,t} - x_{i,j,t-1})^+$.

F. DAHM problem formulation

The problem can be summarized as follows:

Given an application with specific delay requirement, a cloud in which the application can be dynamically hosted, the spatio-temporal variation of electricity price, the spatio-temporal variation of the number of online users, what would be the cost-efficient hosting of the application in the cloud over time?

The cost includes energy cost, performance violation cost and possible migration cost. All aforementioned costs are assumed to be monetary. Therefore, we can model the application hosting problem as an optimization problem where the objective is minimizing the total cost as follows:

minimize

$$\text{Cost} = \text{Cost}^{\text{energy}} + \text{Cost}^{\text{perf. viol.}} + \text{Cost}^{\text{migration}} \\ = \left(\sum_{j=1}^{|A|} x_{i,j,t} n_{j,t} c_i p_{i,t}^{util} + n_{i,t}^{VM} p_i^{idle} \right) PUE_i e_{i,t} \tau \\ + \textstyle\sum_{i=1}^{|S|} \alpha_i (n_{i,t}^{VM} - n_{i,t-1}^{VM})^+ \\ + \textstyle\sum_{t=1}^{T} \displaystyle\sum_{i=1}^{|S|} \sum_{j=1}^{|A|} \eta n_{j,t} x_{i,j,t} (d_{i,j,t} - d^{ref})^+ \\ + \beta \sum_{t=1}^{T} \sum_{i=1}^{|S|} \sum_{j=1}^{|A|} (x_{i,j,t} - x_{i,j,t-1})^+ \tag{2}$$

subject to

$$\forall i, j, t: \quad x_{i,j,t} \in \{0,1\} \text{ and, } \sum_{i=1}^{|S|} x_{i,j,t} \geq \tfrac{n_{j,t}}{L}, \tag{3}$$

$$\forall i, j, t: \quad n_{i,t}^{VM} \in \mathbb{N} \text{ and, } n_{i,t}^{VM} \geq \frac{\sum_{j=1}^{|A|} x_{i,j,t} n_{j,t}}{n_i^{th}}, \tag{4}$$

$$\forall i, t: \quad n_{i,t}^{VM} \leq s_{i,t}. \tag{5}$$

In the above formulation, L denotes a very large number such that $\frac{n_{j,t}}{L} \leq 1, \forall j, t$. Cost minimization is subject to the following constraints:

- *Service constraint* (Eq. 3), which asserts that users of every area are only assigned to one data center, i.e., since $x_{i,j,t}$ is a binary variable, an area whose number of users are greater than zero is assigned to a data center;
- *Idle power constraint* (Eq. 4), which makes sure that idle power consumption is counted per number of VMs assigned to the application; and
- *Capacity constraint* (Eq. 5), i.e., the number of assigned VMs to an application in a data center should not exceed the available VMs (represented by $s_{i,t}$) in the data center.

The solution of this problem specifies, at each time, how many VMs in each data center should be assigned to an application (i.e., $n_{i,t}^{VM}$) and which areas should be assigned to which data centers (i.e., $x_{i,j,t}$). Observe that some of the variables in the problem are binary (i.e., $x_{i,j,t}$) and some are integers (i.e., $n_{i,t}^{VM}$). Therefore, due to linearity of all equations, the problem is a Mixed Integer Programming, MIP. Further, the online form of the problem is reminiscent of the well-known Metrical Task System (MTS), where the requests can only be served by a single server. DAHM however makes the problem complex since requests can originate from multiple areas at the same time and the application can be hosted at multiple data centers.

Under non-zero migration cost (i.e., $\beta > 0$), an optimal solution for DAHM can only be provided if information about electricity price, population and distribution of online users, and the capacity of data centers are all available in advance. However, an online algorithm needs to be designed for practical purposes. The following section presents the online algorithms for solving DAHM.

IV. Solutions of DAHM

Each presented algorithm below solves the problem at the beginning of each epoch, t, based on the current hosting state (i.e., $x_{i,j,t-1}$) of the application, and the electricity price ($e_{i,t}$) and traffic behavior (i.e., distribution and population of online users) of the upcoming epoch. The daily history of electricity price can be used to determine the electricity price (since the daily electricity price does not change a lot). Also, due to cyclic behavior of online users the traffic behavior is predictable.

A. OnlineMIP algorithm

The online version of the DAHM problem (Eq. 2), i.e., without summation over time in all the terms in Eq. 2, is solved using a MIP solver, *GLPK* [19].

B. OnlineGreedy algorithm

At the beginning of an epoch, t, this algorithm associates a cost metric for each pair of data center i and area j, which corresponds to the cost model in the objective function (i.e., Eq. 2). The algorithm then assigns area j to the data center i (i.e., creates a new VM, migrates VM from a different data center assigned for area j in previous epoch to data center i, or uses the existing VMs of data center i to serve requests from area j) where the $cost_{i,j,t}$ is minimum over all data centers. Therefore, the time complexity for this algorithm is linear in number of areas and data centers, i.e., $O(|A||S|)$, which is lower than MIP that is exponential at the worst case.

C. OnlineCOB, Cost oblivious algorithm

We use conventional performance oriented load balancing assignment as a baseline algorithm to evaluate the cost efficiency of our approach. In this approach, (i) each area is assigned to a data center whose delay is the least among all other data centers, (ii) load is balanced among data centers whose delay with respect to areas are the same. This is the approach that is currently used for mirror severs [26]. Also, the number of VMs at each data center is dynamically adjusted at each epoch according to the size of incoming traffic. This algorithm is referred as Online Cost OBlivious , *OnlineCOB* in the rest of the paper.

Table III: Data centers' characteristics.

DC	Elec. price model	peak power(W)	PUE	case*
DC1	Mountain View, CA.	320	1.3	homo
DC1	Mountain View, CA.	400	1.5	hetro
DC2	Houston, TX	320	1.3	both
DC3	Atlanta, GA	320	1.3	both

*: The characteristics of DCs for the homogeneous and heterogeneous case study.

V. Simulation Study

A. Simulation Setup

We simulate a cloud consisting of three data centers, where characteristics of data centers are set according to realistic data. To this end, we assume data centers are located at following three locations: Atlanta, GA; Houston, TX; and Mountain View, CA. namely DC1, DC2 and DC3, respectively. These locations correspond to the location of three major Google data centers. We used the actual electricity price for the above locations [6] (see Fig. 2). Note that, in reality, each data center provider may have different electricity price contracts, i.e., lower electricity price than households. However, the electricity cost can be defined according to actual electricity price or the type of energy source (green or brown). Therefore, the electricity price of Fig. 2 is used as an *example* to show the cost saving benefit of DAHM to leverage electricity cost. The coverage area of the three data centers is chosen based on the distance between the US states and the data centers.

1) Data center types: Three homogeneous data centers are considered for the simulation with state-of-the-art servers (e.g., IBM Systems x3650 M2: idle power 100 and peak power 320 watt) and very low PUE (we use 1.3, which is the PUE of the state of the art [9]). However, to show the efficiency of DAHM solution under different energy proportionality of servers, the Idle to Peak power Ratio (IPR) [14] of servers is changed between zero (ideally energy-proportional server) and 0.6 (old servers). The maximum number of servers for each data center is set to 25.

To model the utilization of servers, we assume that each online user imposes 0.00005 utilization to each server (i.e., $c = 0.0005$) and that each server can handle at most 2000 user requests per second. The server utilization thresholds, u_i^{th}, are set to 75%[1]. The d^{ref} is set to 150 ms, and data centers' reference delay, d'^{ref} is set to 6 ms [27]. Also, for the sake of simplicity, we assume each VM occupies one physical server.

2) Workload distribution: We used one day (March 17 2011) workload trace of an entertainment Web site hosted at *Go Daddy*. Using *Google Analytics*, we collected the hourly total number of visitors to the Web site from different USA states (see Fig. 4). USA states

[1] This value was determined from anecdotal Web searching. It does not affect the validity of the results but only the amount of savings.

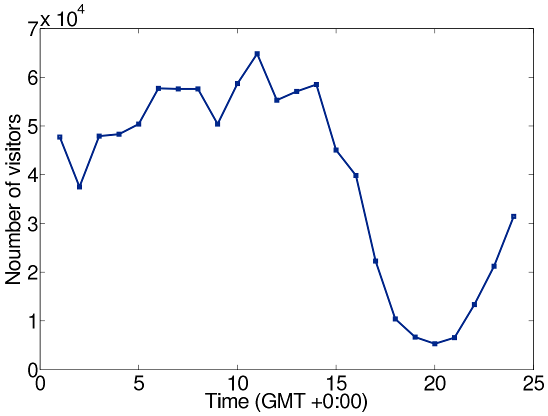

Figure 4: Hourly number of U.S. online users for an entertainment web site hosted at Go Daddy on 17th March, 2011.

Table IV: Tradeoff between performance (delay) violation and total cost saving of DAHM compared to COB-online under different performance violation cost.

Algorithm	$\eta = \eta_0$ saving	viol.	$\eta = 2\eta_0$ saving	viol.	$\eta = 10\eta_0$ saving	viol.
MIP-online	13-14%	0.1-7%	12-13%	0-0.5%	7-7.5%	0

The value of η_0 is set to 0.000001.
The saving values are given in a range from IPR=0 to IPR=0.6.

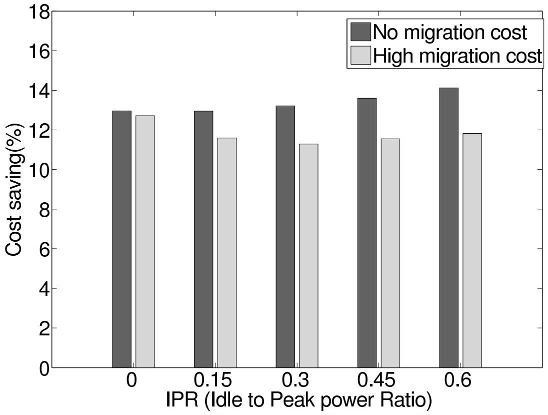

Figure 5: The percentile cost saving of OnlineMIP with respect to OnlineCOB under different IPR of servers (homogeneous data centers).

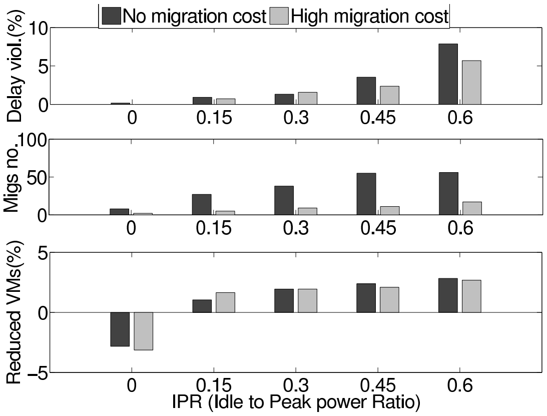

Figure 6: The performance of OnlineMIP with respect to OnlineCOB under different IPR of servers (homogeneous data centers).

are categorized under different areas depending on: (i) delay with respect to data centers, (ii) time zone (for predictability of traffic), and (iii) the desired flexibility of workload distribution. Having small areas (i.e. large number of areas) there are more flexibility for distribution of workload among the data centers. However, the time complexity of the algorithm increases. In our simulation, we used twenty areas over the U.S. The workload is scaled up to the data centers capacity.

3) Experiments performed: We performed different experiments to show the cost saving of DAHM under different of energy proportionality of servers, migration cost and heterogeneity of data centers cases. To this end, experiments are performed under two cases of homogeneous data centers, where all data centers have the same power efficiency (except for the electricity price), and heterogeneous data centers where one data center has different power efficiency compared to two other data centers (see Table III). We also provide results for both high migration cost and zero migration cost cases that are associated with stateful and stateless applications, respectively. We used *GLPK* solver under MATLAB 2009, to run the offline solution (Eq. 2) and OnlineMIP solution.

B. Energy proportionality of servers

The DAHM cost saving under different Idle to Peak power Ratio, (IPR) of servers, shown in Fig. 5, interestingly indicates that DAHM's cost saving increases by increasing IPR under zero migration cost. This cost saving is due to both leveraging the variation of electricity price and the minimization of number of required VMs across all data centers. The results show the benefit of DAHM over previous models, where cost saving was maximized using ideally energy-proportional servers and decreased significantly when servers have IPR of greater than zero [18], [17].

Fig. 6 shows that when servers are assumed to be ideally energy-proportional (IPR=0), DAHM increases the number of VMs to leverage electricity price variation. However, under non-zero IPR and zero migration cost, DAHM always decreases total the number of VMs by 1-3% with respect to COB-online. Fig. 6 also shows that DAHM solution sometimes violates the delay. The delay violation is calculated as a percentage of total users whose delay is more than d^{ref}. The delay violation is decreased by increasing the SLA violation cost. The results in Table IV show that by doubling η, the cost saving of DAHM under IPR=0.6 decreases by one

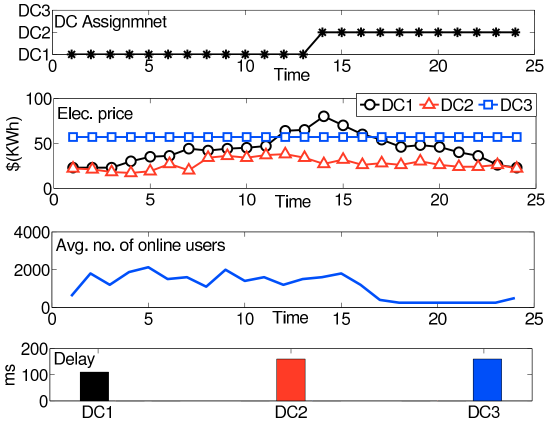

Figure 7: The data center host and workload density of an area over time (homogeneous data centers).

percent, and delay violation is decreased to 0.5 percent.

Under high migration cost, reduction of number of VMs over cloud is observed to be less than the zero migration cost case (see Fig. 6). This is the reason why DAHM's cost saving for high migration cost case marginally decreases compared to zero migration cost case, and that its cost saving marginally decreases by increasing IPR (12% cost saving for IPR=0, and 11.2% for IPR=0.6).

The cost saving of DAHM over IPR under migration cost decreases for IPR=0 up to IPR=0.3 and then increases. The reason is that for low IPR, cost saving due to consolidation (i.e., reducing number of VMs) is low, where DAHM prefers higher number of VMs to avoid migration cost. In other words, in this case, DAHM mostly performs workload management whose cost saving decreases for larger IPRs. For large IPRs (IPR=0.3 upto IPR=0.6), however, the cost-saving benefit of consolidating is high, which causes DAHM to reduce the idle power by increasing the consolidation rate. Hence, the cost-saving of DAHM is higher at larger IPRs.

C. Leveraging workload variation

The results in the previous section show that DAHM solution yields cost saving by migrating VMs over cloud. The results in Fig. 7 show that migration is tightly correlated with the electricity cost, workload density, and the delay constraint. It can be seen in the figure that when workload density of the tested area is low, and the electricity price of the data center where it is assigned increases, DAHM changes its data center host to the Data center 2. However, this causes the delay requirement of a few users to be marginally violated.

D. Leveraging heterogeneity of data centers

In practice, data centers are not identical and have different energy efficiency which can be leveraged for cost-efficient application hosting. To investigate the potential saving of heterogeneous data centers and its contribution in the total cost efficiency of DAHM, we make DC1 to be less energy-efficient than data center number two and three. To this end, the PUE of DC1 is changed to 1.5 and the peak power of servers is changed to 400 (see Table III). The results in Fig. 8 show that DAHM cost saving increases from 14% for the homogeneous data center case to 20% for the heterogeneous data center case. In this case, even if the migration cost is high, cost saving of DAHM increases with increase in servers' IPR. This saving comes from minimizing the number of VMs over cloud to leverage the electricity cost and energy efficiency of data centers (see Fig. 9). Note that the delay violation of the DAHM under heterogeneous data center case, shown in Fig. 9 is not increased compared to the homogeneous data center case.

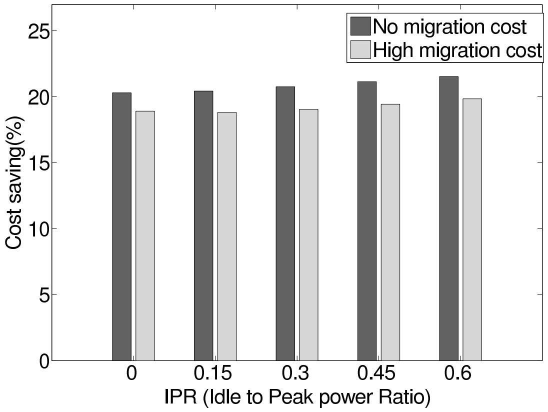

Figure 8: The percentile cost saving of OnlineMIP with respect to OnlineCOB under different IPR of servers (heterogeneous data centers).

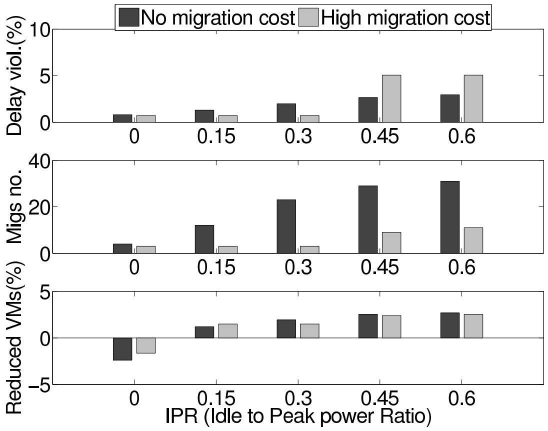

Figure 9: The performance of OnlineMIP with respect to OnlineCOB under different IPR of servers (heterogeneous data centers).

E. Offline and Online solutions

Under zero migration cost, OnlineMIP provides optimal solution for DAHM problem. In all of our experi-

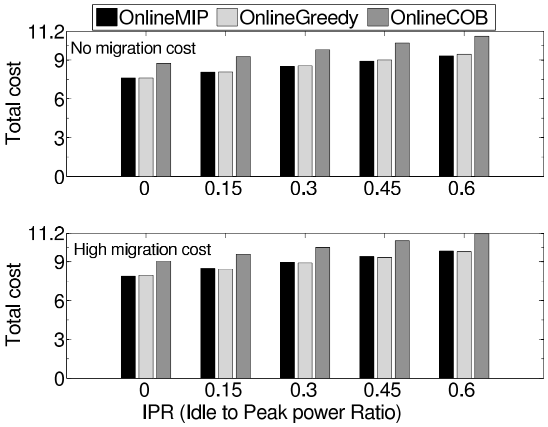

Figure 10: Total cost of several DAHM solution under no and high migration cost.

ments, all DAHM solution including OnlineMIP, take a fraction of second to solve DAHM for every epoch, all running on a 2.8 GHz Intel Pentium system. However, MIP in the worse case may take exponential time to complete. For that reason we developed OnlineGreedy algorithm. Results shown in Fig. 10, indicates that OnlineGreedy algorithm can competitively save cost compared to OnlineMIP under zero migration cost. OnlineMIP has marginal saving of less than 1% compared to OnlineGreedy.

Under high migration cost, neither OnlineMIP nor OnlineGreedy can provide optimal solution. As shown in Fig. 10, the cost saving of OnlineMIP and OnlineGreedy are almost the same in this case. The optimal solution under high migration cost can only be achieved offline. Comparison of DAHM offline-optimal with respect to online solutions [2] indicates that offline optimal always achieves higher cost savings, upto 0.5%, compared to online solutions (its cost saving increases by increase in number of epochs). Online algorithm with competitive bounds is left as future work.

F. Issues of implementing DAHM in practice

In our simulation study, we assume that in the beginning of each epoch the data for workload and electricity price is available, but in practice these data should be predicted. Both workload and electricity are predictable, however the prediction error may decrease the saving marginally.

In practice, the network delay among end users and data centers vary over time and increases in the peak traffic time. DAHM is sensitive to network delay as well as data center delay. However, in the simulation study, we consider the network delay to be a constant.

For the sake of simplicity we assumed a constant cost per migration. Incorporating migration cost as a function of number of users is a trivial task. DAHM can be considered as a central controller and should be frequently updated with data of network delay, electricity price and history of workload from data centers. Since these data should be sent at each epoch (half an hour to several hours), its overhead is negligible. Therefore, DAHM can be implemented in practice under low overhead and yield cost savings.

These practical issues will be tested and examined using the BlueTool research infrastructure [8], [28], which offers a small data center for experimentation with innovative management schemes such as DAHM.

VI. Conclusions

This paper presents problem formulation and algorithms for DAHM, which allow cloud providers to host Web allocation cost efficiently in a dynamic fashion. The problem is formulated according to a cost model that accounts for energy cost of data centers, delay requirement, and traffic behavior of applications as well as live migrations. We design low complexity solutions to the problem and perform simulation study under realistic data and make the following conclusions: (i) dynamic hosting minimizes the total number of VMs over cloud and yields significant cost savings by removing idle power cost; (ii) VM migration can leverage the temporal and spatial variation of electricity price, workload and data centers' energy efficiency to minimize total cost, and (iii) cost value of SLA violation can be controlled to meet performance goals and still achieve cost savings. Developing online algorithms with a low competitive ratio is left for future work.

Lastly, this work conventionally considers stationary users that "log in" or out from the service without changing location. Future work should consider mobile users who need continuous service from the cloud. Such consideration would require designing the dynamic hosting scheme along the lines of a registration area scheme with dynamically overlapping registration areas [29].

References

[1] P. B. Beskow, K. H. Vik, P. Halvorsen, and C. Griwodz, "The partial migration of game state and dynamic server selection to reduce latency," *Multimedia Tools and Applications*, vol. 45(3), 2009.

[2] K. Kumar and Y.-H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *Computer*, vol. 99, pp. 51–56, 2010.

[3] J. G. Koomey, C. Belady, M. Patterson, A. Santos, and K.-D. Lange, "Assessing trends over time in performance, costs, and energy use for servers," Microsoft Corp. and Intel Corp., Tech. Rep., 2009.

[2] Due to high time complexity of the offline optimal algorithm, we just ran the algorithm for few hours instead of entire 24 hours.

[4] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A.Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A Berkeley view of cloud computing," in *Technical Report No. UCB/EECS-2009-28,*, University of California at Berkley, USA, 2009.

[5] S. K. S. Gupta, T. Mukherjee, G. Varsamopoulos, and A. Banerjee, "Research directions in energy-sustainable cyber-physical systems," *Elsevier Comnets Special Issue in Sustainable Computing (SUSCOM), Invited paper*, vol. 1, no. 1, pp. 57–74, 2011.

[6] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing electricity cost: optimization of distributed Internet data centers in a multi-electricity-market environment," in *Proc. IEEE INFOCOM*, 2010, pp. 1–9.

[7] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos, "Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 11, pp. 1458–1472, 2008.

[8] S. K. S. Gupta, R. R. Gilbert, A. Banerjee, Z. Abbasi, T. Mukherjee, and G. Varsamopoulos, "GDCSim: A tool for analyzing green data center design and resource management techniques," in *Proc. of International Green Computing Conference(IGCC11)*. IEEE, 2011.

[9] "Quick start guide to increase data center energy efficiency," General Services Administration (GSA) and the Federal Energy Management Program (FEMP)., Tech. Rep., 2010.

[10] http://gigaom.com/2008/10/01/google-data-centers-more-efficient-than-the-industry-average/.

[11] A. Krioukov, P. Mohan, S. Alspaugh, L. Keys, D. Culler, and R. Katz, "NapSAC: design and implementation of a power-proportional web cluster," in *Proc.of the first SIGCOMM workshop on Green networking*. ACM, 2010, pp. 15–22.

[12] B. G. Chun, G. Iannaccone, G. Iannaccone, R. Katz, G. Lee, and L. Niccolini, "An energy case for hybrid datacenters," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 1, pp. 76–80, 2010.

[13] D. Tsirogiannis, S. Harizopoulos, and M. A. Shah, "Analyzing the energy efficiency of a database server," in *Proc. of the 2010 international conference on Management of data SIGMOD'10*. ACM, 2010, pp. 231–242.

[14] G. Varsamopoulos, Z. Abbasi, and S. K. S. Gupta, "Trends and effects of energy proportionality on server provisioning in data centers," in *International Conference on High performance Computing Conference (HiPC2010)*, 2010, pp. 1–11.

[15] D. Meisner, B. Gold, and T. Wenisch, "PowerNap: eliminating server idle power," *ACM SIGPLAN Notices*, vol. 44, no. 3, pp. 205–216, 2009.

[16] B. Guenter, N. Jain, and C. Williams, "Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning," in *Proc. IEEE INFOCOM, Shanghai, China*, 2011, pp. 702–710.

[17] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for Internet-scale systems," in *Proc. ACM SIGCOMM*, 2009, pp. 123–134.

[18] K. Le, O. Bilgir, R. Bianchini, M. Martonosi, and T. Nguyen, "Managing the cost, energy consumption, and carbon footprint of Internet services," *SIGMETRICS Perform. Eval. Rev.*, vol. 38, no. 1, pp. 357–358, 2010.

[19] [Online]. Available: http://www.gnu.org/software/glpk/

[20] L. Rao, X. Liu, M. Ilic, and J. Liu, "MEC-IDC: joint load balancing and power control for distributed Internet data centers," in *Proc. of the 1st ACM/IEEE International Conference on Cyber-Physical Systems*, 2010, pp. 188–197.

[21] Y. Zhang, Y. Wang, and X. Wang, "Capping the electricity cost of cloud-scale data centers with impacts on power markets," in *Proc. of the 20th international symposium on High performance distributed computing*. ACM, 2011, pp. 271–272.

[22] J. Chase, D. Anderson, P. Thakar, A. Vahdat, and R. Doyle, "Managing energy and server resources in hosting centers," in *Proc. of the eighteenth ACM symposium on Operating systems principles (SOSP 01)*, 2001, pp. 103–116.

[23] Z. Abbasi, G. Varsamopoulos, and S. K. S. Gupta, "Thermal aware server provisioning and workload distribution for Internet data centers," in *ACM International Symposium on High Performance Distributed Computing (HPDC10)*, 2010, pp. 130–141.

[24] D. Kusic, J. O. Kephart, J. E. Hanson, N. Kandasamy, and G. Jiang, "Power and performance management of virtualized computing environments via lookahead control," *Cluster Computing*, vol. 12, pp. 1–15, 2009.

[25] M. Lin, A. Wierman, L. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," in *Proc. IEEE INFOCOM*, Shanghai, China, 2011, pp. 10–15.

[26] M. Emens, D. Ford, R. Kraft, and G. Tewari, "Method of automatically selecting a mirror server for web-based client-host interaction," 2003, patent 6,606,643.

[27] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, and N. Gautam, "Managing server energy and operational costs in hosting centers," *SIGMETRICS Performance Evaluation Review*, vol. 33, no. 1, pp. 303–314, 2005.

[28] S. K. S. Gupta, G. Varsamopoulos, A. Haywood, P. Phelan, and T. Mukherjee, *Handbook of Energy-Aware and Green Computing*. Chapman and Hall/CRC, 2012, no. 45, ch. BlueTool: Using a computing systems research infrastructure tool to design and test green and sustainable data centers.

[29] G. Varsamopoulos and S. K. S. Gupta, "Dynamically adapting registration areas to user mobility and call patterns for efficient location management in pcs networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 5, pp. 837–850, 2004.