Cloud Data Centers Revenue Maximization using Server Consolidation: Modeling and Evaluation

Mohammad Wardat
Iowa State University
IA, USA

Mahmoud Al-Ayyoub and Yaser Jararweh
Jordan University of Science and Technology
Irbid, Jordan

Abdallah A. Khreishah
New Jersey Institute of Technology
NJ, USA

***Abstract*—Cloud based data centers (DC) offering of virtually unlimited processing and storage capabilities faced a critical downside related to the enormous power consumption requirements. Powering and cooling DC has considerable share of the total DC operational costs in addition to the carbon tax. Server consolidation has been proposed as one viable solution that guarantees efficient power usage while maintaining quality of service requirements for the DC's services. This approach is simply about turning on and off the DC resources based on the incoming workload. In this paper, we are extending our previous work on expansion strategies for cloud services providers by considering the server consolidation technique in order to maximize the DC revenues while maintaining their service quality and reducing their carbon footprints.**

I. Introduction

Cloud computing (CC) paradigm promised services to be provided to the end users with the minimum users awareness of the hardware/softwares infrastructure running their applications. This will enable the users to focus on their core applications development and their business model enhancements. The burden of handling the logistics related to the computing infrastructure, networking, storage and security is moved to the cloud service providers (CSP) on a pay-as-you-go model. The increasing trend of companies and individuals migrating to CC causes a similar increase in the number of CSP. CSP are competing with each other to provide the best Quality of Service (QoS) to their customers while maximizing their revenues [1], [2].

To achieve the aforementioned objectives, the CSP are forced to expand their services capabilities and to build more and more geographically distributed data centers (DC) to handle the increasing demands. At the same time, they are optimizing their DC activities to reduce their operational costs. Power consumptions used in operating and cooling DC represent a major chunk of the total operational costs for running a cloud DC. Hence, optimizing power consumptions in DC received a lot of attention from both academia and the industry.

Revenue maximization for CSP requires a holistic view of DC expansion strategies and operational costs minimization (e.g., power usage). The DC expansion strategies focus on creating new DC to handle the ever increasing demand for resources while maintaining high QoS provided to the users based on the Service Level agreement (SLA) with them. This needs to be in line with the return on investment for the CSP. On the other hand, the power consumption optimization will consider the economical aspects of the DC operations that will satisfy the incoming workload. While expanding current DC or building new DC depend on different factors like land costs and the availability of the supporting infrastructure (e.g., the electrical power sources), the power consumption optimizations can be implemented on a fine grained level like computing servers or network switches.

In this paper, we are extending our previous work on the expansion strategies for cloud DC, which includes building new DC or expanding currently operating DC [1], [2]. The extension includes the power usage optimization using an approach known as server consolidation. Current servers are powered on all the time with average server utilization of about 15% to 20%. The server consolidation aims to increase the average utilization by reducing the number of active servers. This is achieved by powering off underutilized servers without impacting the ability of the system to satisfy the customers' QoS requirements and meet the SLAs. Our proposed model takes into consideration the potential availability of renewable energy sources to power the DC.

The rest of this paper is organized as follows. A detailed related work is presented in Section II. In Section III, we present our extended system model with all constraints. The model evaluation and the results discussion are presented in Section IV. Finally, we conclude this paper in Section V.

II. Related Works

Perhaps, the closest work to ours is [3]. The authors employed server consolidation where server workload predictions were used to make decisions in order to efficiently manage the system and minimize energy consumption. They suggested a new model for finite buffer which allows the server to be turned on/off by a load-dependent control using multi-parallel hysteresis thresholds.

In [4], the authors presented a snapshot-based solution for the server consolidation problem in order to reduce energy consumption. They also took into account some issues such as reducing the total number of virtual machine (VM) migrations (where VM are migrated to switch hosts off to reduce energy consumption) and consolidating the save loads of running server.

In [5], the authors proposed a new algorithm called CUBE-FIT, which can use a fewer number of server to handle the

978-1-5386-5979-3/18/$31.00 ©2018 IEEE

workloads. Also, the algorithm can handle the system when multiple server failures while reducing the cost and ensuring that no server becomes overloaded.

Recently, optimizing power cost (instead of consumption) has become the focus of many interesting works. Some of these works have utilized the server consolidation technique in order to achieve their goals. In general, the revenue maximization problem does not appear to have been studied extensively before. Below, we discuss the most relevant papers to our work.

Electricity prices in the wholesale market vary a lot from time to time and from region to region. This means that different DC will have different costs to do the same job. The same thing applies to the same DC doing the same job at different times. Accordingly, the authors of [6] designed a DC load placement strategy to optimize the power cost by employing the DC that is located at cheap price regions within an "acceptable" distance from the client.

In [7], the authors proposed a technique for the conservation of energy in a heterogeneous cluster of workstations. The technique minimizes the power consumption by turning cluster nodes on to handle the workload efficiently and switching the idle node off to save power under lighter load. They proposed an algorithm that runs on a master node to monitor the load on resources (disk storage, CPU and network interface) and makes decisions about switching nodes on/off to minimize the power consumption while taking into account SLA and QoS. The algorithm runs at two levels: (1) at the operating system level for cluster cycle servers and (2) at the application level for a cluster-based, locality conscious network server. The proposed approach can be applied to multi-application mixed-workload environments with fixed SLA.

In [8], the authors proposed a power consolidation approach for VM placement in virtualized heterogeneous computing environments by taking the advantage of min-max and shares parameters of VM. The techniques allow the DC to allocate resources in order to run heterogeneous applications based on available resources, power costs, and application utilities.

In [9], the authors investigated and formulated problem for poweraware application placement in the environment with heterogeneous virtualized server clusters by taking into account the migration and power costs. They divided the problem into two parts. In the first part, they presented multiple ways to capture the cost-aware application placement problem, while, in the second part, they presented the pMapper architecture and placement algorithm to solve one practical formulation of the problem, which is to minimize power consumption and maximize performance.

Recent studied focused on the dynamic voltage/frequency scaling for reducing the power consumption in clusters and DC. An interesting work was introduced in [10] to provide a theoretical queuing model to predict the optimal power allocation in virtualized heterogeneous server farm while minimizing the mean response time. The idea is to find the specific relationship between frequency and power for optimal power allocation at the level of server farms.

The CPU, disk storage and network interface are the main components in the DC that consume the power. In comparison to other resources in the DC, CPU consume large amounts of energy. For this reason, many recent studies [10], [11] focused on managing its power consumption and efficient usage. Some of these studies showed that an idle server consumes approximately 70% of the power consumed when the server running at full speed. So, they proposed many techniques, such as Dynamic Voltage and Frequency Scaling (DVFS) on CPU, for switching the idle servers off in order to reduce power consumption.

The literature related to DC demand response (DR) is of concern to diverse groups in many areas. With the increasing VM density in cloud DC, it is becoming more difficult to manage the physical resources. In [12], the authors designed and evaluated revenue driven dynamic resource allocation to achieve the objective of maximizing SLA-constrained revenue. The authors integrated the performance models with a hill-climbing algorithm to achieve the objective.

At present, many research efforts have been devoted to increasing efficiency of DC by reducing power consumption and carbon dioxide emissions. The goal of these efforts is to efficiently utilize the available resources and to reduce energy consumption and thermal cooling costs. In [13], the authors proposed an effective dynamic scheduler to maximize the application throughput and minimize the computing-plus-communication energy consumption. The overall goal is to reduce the energy consumption while guaranteeing high QoS in cloud DC.

CSP face the major challenge of the power demand in DC on a more scalable curve because the growing popularity of Cloud applications. To minimize the power consumption and to reduce economical and environmental impact, it is important to understand the relationship between power, DVFS and consolidation. In [14], the authors proposed a DVFS policy that reduces power consumption while preventing performance degradation, and a DVFS-aware consolidation policy that optimizes consumption. Because the CSP are bound to strict SLA conditions, the authors took into account the DVFS configuration that would be necessary to maintain QoS. In [15], the authors developed framework to address frequency regulation by controlling facility energy consumption via battery charging/discharging with no performance impact on the workloads.

Another study [16] gave attention to designing hardware and algorithms that can adapt energy usage by focusing on speed-scaling and take into consideration the power-capping. In [16], the authors proposed a dynamic server power capping technique to manage and focus on how a DC should respond to the energy price in order to reduce its energy bill while maintaining the desired QoS. Also, the authors constructed and solved an optimization problem With the goal of optimizing the average power consumption and regulation quantity in order to minimize the cost of energy.

Recently, the DR programs have been the focus of many researchers to significantly reduce peak demands and allow

for easier integration of renewable energy into the grid. One field of DR has become increasingly important with increasing energy consumption is the coincident peak pricing programs. In [17], the authors developed two algorithms for the DC that participate in DR programs by combining the workload shifting of interactive/batch workload and the use of local power generation in order to avoid coincident peaks and, thus, reducing the energy expenditure. Then, the peak power usage fee can be effectively mitigated. Thus, participating in DR programs can reduce electricity cost significantly; however, this comes at the expense of server availability.

The workload scheduling for taming DC peak power demand charge has been studied in [18]. The authors designed a hierarchical framework for optimizing DC electric utility bills via both partial execution and workload deferring mechanisms, where the workload deferring can be adopted if the workload is latency-tolerable, e.g., for batch workload. Also, the authors have investigated server resource management techniques (e.g., switching off idle servers) to modulate energy demand. There are clearly many important research questions to address, compared with the previously mentioned related works. It is clear that many significant challenges are yet to be addressed.

The importance of the proposed work comes mainly from addressing the revenue maximization problem by employing server consolidation in DC. Our model assigns incoming tasks to the minimum number of active servers in the DC and turns off unused servers while taking into account the trade-off between maximizing the revenue and minimizing the operational cost of the DC. This model is extending the formulation of the DC expansion problem by using a mixed integer linear programming approach.

III. System Model

In this section, we extend our previous work [1], [2] by employing the server consolidation technique in order to face the increasing demands and optimize the revenue. We formulate this optimization problem using mixed integer linear programming (MILP).

Following the notation of [1], [2], we denote the set of users locations by U and the set of DC locations by S (including current locations as well as candidate locations on which new DC can be built). To denote whether a DC is built on location s or not, we use the set of binary variables, $X = \{x_s | s \in S\}$. We use L_u^h to represent the total number of service requests (SR) originating from users on location u at time h. We also use $\lambda_{s,u}^h$ to represent the fraction of L_u^h that is serviced by the DC s. Finally, we define $\lambda = \{\lambda_{s,u}^h | s \in S, u \in U, h \in H\}$ and enforce the following constraint in order to ensure that no request is denied.

$$\sum_{s \in S} \lambda_{s,u}^h = L_u^h, \quad \forall h \in H \tag{1}$$

The ability of a DC s to handle SR from users on location u at time h is represented by the binary variable $y_{s,u}^h$. Let $Y = \{y_{s,u}^h | s \in S, u \in U, h \in H\}$. The following constraint ensures that a DC that is not built yet cannot handle any SR.

$$y_{s,u}^h \leq x_s, \quad \forall s \in S, u \in U, h \in H \tag{2}$$

We also need to ensure that SR are not routed to DC that cannot handle them.

$$0 \leq \lambda_{s,u}^h \leq y_{s,u}^h L_u^h, \quad \forall s \in S, u \in U, h \in H \tag{3}$$

If R_s is the set of servers in DC s, then, the following constraint defines a SR distribution policy to assign SR received by the DC to its servers.

$$\sum_{r \in R_s} \psi_{s,u}^{h,r} = \lambda_{s,u}^h \quad \forall s \in S, h \in H \tag{4}$$

As in [19], we define another binary variable ($I_s^{h,r}$) to represent the operating status of a server r at DC s during time slot h. Let $I = \{I_s^{h,r} | s \in S, r \in R_s, h \in H\}$. Thus, we have the following constraint.

$$I_s^{h,r} \leq x_s, \quad \forall s \in S, r \in R_s, h \in H \tag{5}$$

In order to calculate the power consumption of a server's startup/shutdown process, we use the model of [19]. Let $ESD_s^{h,r}$ and $ESU_s^{h,r}$ be the power consumption of server r during startup and shutdown delays, respectively. The startup variable, $UP_s^{h,r}$, for server r is set to 1 once server r is invoked at time slot h in DC s; otherwise, it is set to 0. We also define a shutdown variable $DN_s^{h,r}$, which will be set to 1 when server r is indicated to shutdown at time slot h in DC s and 0 otherwise.

To calculate the power consumption of each server r at DC s at time slot h, we employ [20]'s model.

$$\begin{aligned} P_s^{h,r} &= \Big[I_s^{h,r}(P_{\text{idle}} + (E_{\text{usage}} - 1)P_{\text{peak}}) \\ &+ I_s^{h,r}(P_{\text{peak}} - P_{\text{idle}})\gamma_s^h + x^s \epsilon \Big] \\ &+ ESU_s^{h,r} \times UP_s^{h,r} + ESD_s^{h,r} \times DN_s^{h,r} \end{aligned} \tag{6}$$

$$\gamma_s^h = \frac{\sum_{u \in U} \psi_{s,u}^{h,r}}{\mu_s^r} \times I_s^{h,r} \tag{7}$$

$$\sum_{r \in R} P_s^{h,r} \leq P_s^{h,\max}, \quad \forall s \in S, h \in H \tag{8}$$

$$2D_{s,u} y_{s,u}^h \leq D^{\max}, \quad \forall s \in S, u \in U, h \in H \tag{9}$$

$$\gamma_s^{t,h} \leq \gamma^{\max}, \quad \forall s \in S, h \in H \tag{10}$$

As in [19]. We define the Constraints 11–22. To represent the initial conditions for each server when it has already been started, during waking up process, or shutdown, Constraints 11-14 are used. For "startup delay" periods, servers are assumed to be in the OFF status. When a server is initially at waking up process, its status is turned to ON status for

"minimum ON" time. A server should stay in the OFF status for "shutdown delay" plus "minimum OFF" time when it is initially at shutdown status. If the server is initially ON, the following holds.

$$I_s^{h,r} = 1, h \in [1, MU_s^r - TUO_s^r] \tag{11}$$

If the server is initially during WAKEUP process, the following holds.

$$I_s^{h,r} = 0, h \in [1, DSU_s^r - TWO_s^r] \tag{12}$$

$$I_s^{h,r} = 1, h \in [DSU_s^r - TWO_s^r + 1, DSU_s^r - TWO_s^r + MU_s^r] \tag{13}$$

If the server is initially OFF, the following holds.

$$I_s^{h,r} = 0, h \in [1, MD_s^r + DSD_s^r - TDO_s^r] \tag{14}$$

$$\text{for } h \leq H - DSU_s^r - MU_s^r + 1$$

$$\sum_{n=h+DSU_s^r}^{h+DSU_s^r+MU_s^r-1} I_s^{n,r} \geq MU_s^r \times DN_s^{h,r}$$

Constraints 15-17 represents the "minimum ON" time requirements for each server in each period. Once a server is invoked to wake up, it will stay in the OFF status for "startup delay" periods, and then it will be ON for at least "minimum ON" time.

$$\sum_{n=h}^{h+DSU_s^r-1} I_s^{n,r} \leq DSU_s^r \times (1 - UP_s^{h,r}) \tag{15}$$

$$\text{for } H - DSU_s^r - MU_s^r + 2 \leq h \leq H - DSU_s^r$$

$$\sum_{n=h+DSU_s^r}^{H} I_s^{n,r} \geq (H - h - DSU_s^r + 1)(UP_s^{h,r})$$

$$\sum_{n=h}^{h+DSU_s^r-1} I_s^{n,r} \leq (DSU_s^r)(1 - UP_s^{h,r}) \tag{16}$$

$$\text{for } H - DSU_s^r + 1 \leq h \leq H$$

$$\sum_{n=h}^{H} I_s^{n,r} \leq (H - h + 1)(1 - UP_s^{h,r}) \tag{17}$$

$$\text{for } H \leq H - MD_s^r - DSD_s^r + 1$$

Similarly, constraints 18-19 enforce the "minimum OFF" time requirements for each server in each period. Once a server is instructed to shut down, it will turn to OFF immediately and stay in the OFF status for "shutdown delay" plus "minimum OFF" time.

$$\sum_{n=h}^{h+MD_s^r+DSD_s^r-1} I_s^{n,r} \leq (MD_s^r + DSD_s^r)(1 - DN_s^{h,r}) \tag{18}$$

$$\text{for } H - MD_s^r - DSD_s^r + 2 \leq h \leq H$$

$$\sum_{n=h}^{H} I_s^{n,r} \leq (H - h + 1)(1 - DN_s^{h,r}) \tag{19}$$

Finally, constraint 20 enforces the relationship between a server's startup/shutdown indicator and ON/OFF status. Here, the startup variable $UP_s^{h,r}$ will be equal to 1 once the server is invoked to start up at period h; otherwise, it will be equal to 0. Similarly, the shutdown variable $DN_s^{h,r}$ will be equal to 1 once server is indicated to shutdown at period h and 0 otherwise. Both startup and shutdown indicators are binary variables.

$$I_s^{h,r} - I_s^{h-1,r} = UP_s^{h,r} - DSU_s^r - DN_s^{h,r} \tag{20}$$

In order to reduce the number of binary variables and improve the model efficiency, $UP_s^{h,r}$ and $DN_s^{h,r}$ are modeled as continuous variables by introducing constraints 21-22.

$$UP_s^{h,r} - DSD_s^r \leq 1 - I_s^{h-1,r} \tag{21}$$

$$DN_s^{h,r} \leq I_s^{h-1,r} \tag{22}$$

The objective function is as follows.

$$\begin{array}{cc} \text{Maximize}_{x,m} & \text{RV}(T) - (\text{OPEX}(T) + \text{CAPEX}(T)) \\ \text{Subject to} & \text{Constraints} \quad 1 - 22. \end{array}$$

The overall cost of the DC can be divided into operational cost (OPEX) and capital cost (CAPEX). More formally, CAPEX for a certain year t can be expressed using the following equation.

$$\text{CAPEX}(t) = \sum_{s \in S} (x_s^{t-1} - x_s^t)\text{BC}_s^t + (m_s^{t-1} - m_s^t)\text{SC}_s^t,$$

where BC_s^t represent the cost of building a DC s in year t and $\text{SC}_{t,s}$ represent the cost of buying a server for the DC s in year t.

OPEX for a certain year t can be expressed as follows [21].

$$\text{OPEX(t)} = \sum_{s \in S} \sum_{h \in H} (\theta_s^t P_s^{t,h} + \delta_s^t (\rho_s + 1) P_s^{t,h} + \sum_{u \in U} (\lambda_{s,u}^{t,h} \sigma_{s,u}^t)),$$

where δ_s^t is the carbon tax in location s in year t, ρ_s is the power transmission loss rate location s, $\sigma_{s,u}^t$ is the cost of the bandwidth between user location u and candidate location s and θ_s^t is the price of electricity in candidate location s taken during three different time-of-use price periods: off-peak, mid-peak and on-peak.

Now, the revenue of year t is computed using the following equation [22]: $\text{RV}(t) = ((1 - p(x))\alpha^t \lambda_{s,u}^{t,h} - p(x)\beta^t)$, where $p(x)$ is the probability that the waiting time for a service request exceeds the SLA-deadline, α^t is the service fee that the DC charges the costumers for handling a single service request and β^t is the penalty that the DC must pay for every service request it cannot handle (thus, causing an SLA violation).

A. Renewable Energy

The model discussed so far does not explicitly account for renewable energy, which is one of the biggest concerns related to DC and their effect on the surrounding environment. To address this issue, we reformulate Equation 6 as follows [22].

$$\begin{aligned} P_s^{h,r} &= \Big[\Big[[I_s^{h,r}(P_{\text{idle}} + (E_{\text{usage}} - 1)P_{\text{peak}}) \\ &+ I_s^{h,r}(P_{\text{peak}} - P_{\text{idle}})\gamma_s^h + x^s\epsilon] \\ &+ ESU_s^{h,r} \times UP_s^{h,r} + ESD_s^{h,r} \times DN_s^{h,r}\Big] - x_s G_s^h\Big]^+ \end{aligned}$$

where $[x]^+ = \max\{x, 0\}$ and $G_s^{t,h}$ is the amount of renewable power generated in location s during hour h. The details for the renewable power inclusion is available at [1], [22].

IV. Experiments and Results

Now we will discuss our model implementation and the simulation obtained results.

In our experiment, we will compare this model by using Server Consolidation and previous model in [2]. Also, we set all input variable the same as in [1]. However, we do consider a more realistic case where the number of candidate DC, number of server and initial traffic load as shown in Table I. For the sake of simplicity, we consider the number of servers: 25 OR 50 Servers for each DC. Also, for the traffic load, we choose the total number of SR incoming from all user locations to be between 2.5 and 10 thousand hits/hour. We set the remaining variables as we used in the previous model [1] to compare between the revenue when we apply the server Consolidation and without server Consolidation.

Let's start by experiment 1. In Table I, the initial traffic load is 10000 requests/hour, and the number of server in each candidate DC is 25 servers. From Figure 1, in first year, we see the revenue in server consolidation is positive comparing with another model. The reason that workload less than the peak load and some of servers are consuming power during idle period. This will cause a decrease in revenue, with inflation in the traffic load after 10 years and the workload reach to peak. In this case, all server will be in the peak period and no need to server consolidation in this case, as we see in Figure 1, at year 10, the revenue from the two models are the same.

From Figure 1-6, to analyze the performance improvement in our model, we compare between a model with server consolidation and a model without server consolidation (we assume all server are staying "ON"). We conduct different experiments with different workloads and different numbers of servers.

As we see, the server consolidation returns a better revenue compared with the other model. The difference in revenue between the two models depends on the number of servers in each DC and the traffic load. In each experiment the server consolidation return positive revenue compared with other model that return negative revenue in some case and the reason for that in some case all server will be "ON" and there is no traffic load. In this case the server consume power without any processing. To handle this problem, our proposed model reduces the total electricity consumption in each DC and increases the revenue by considering the severe consideration.

In each experiment there are varying cost reductions. As we see, the server consolidation model is more effective than other model in reducing the electricity special when the number of servers is large in DC and the traffic load is less than the number of servers. This will increase the revenue for cloud provider, and the server consolidation model will help 'to handle request by turning "ON" the require servers and make other servers "OFF". This will save more power when the server idle.

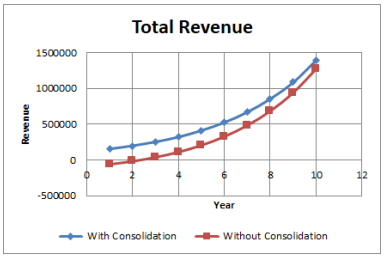

Fig. 1. Experiment Number One

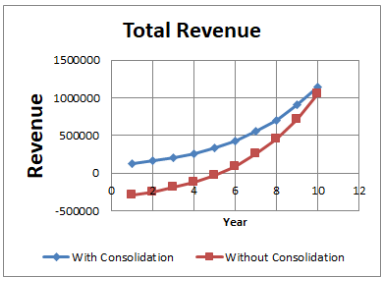

Fig. 2. Experiment Number Two

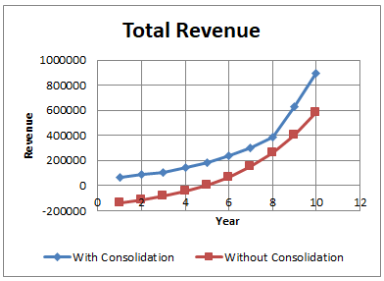

Fig. 3. Experiment Number Three

Now, let us present and discuss the results of the six experiments we conduct. The objective of six experiments are to study how the server consolidation can handle the increasing traffic load and increase the profit for cloud provider. We run our model for 10 years on 7 DC locations, and we use different number of servers and incoming request. in our model we assume that all DC already built and have some servers in it.

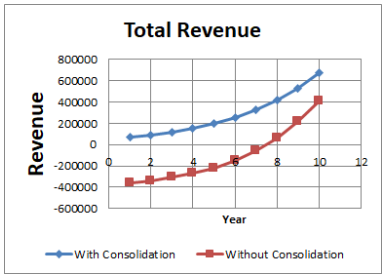

Fig. 4. Experiment Number Four

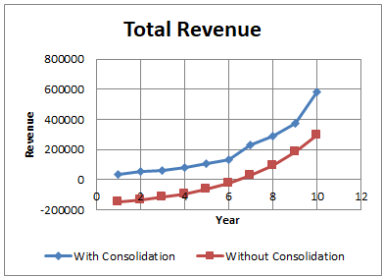

Fig. 5. Experiment Number Five

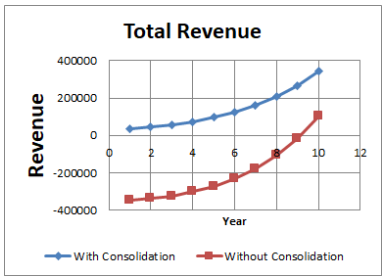

Fig. 6. Experiment Number Six

Since the servers are homogeneous and the maximum number of servers to be placed in a single DC for each experiment shown in Table I.

In the our experiment, we study the effect of using server consolidation and how we can increase the profit. From Figures 1-6, it can be seen that using server consolidation with different number of servers and traffic load for different times of periods generates better annual profits than using previous model in [1]. Moreover, increasing traffic load with using server consolidation also has even more positive effect on the annual profits compare with other model. Finally, increasing number of servers and the number of traffic load whit both settings (with consolidation and without consolidation) causes a big improvement on the annual profits.

TABLE I
EXPERIMENT SETTINGS

	candidate DC	Server	Initial traffic load
experiment 1	7	25	10000
experiment 2	7	50	10000
experiment 3	7	25	5000
experiment 4	7	50	5000
experiment 5	7	25	250
experiment 6	7	50	250

V. CONCLUSION

The increasing demands for cloud services and the efforts of cloud service providers to fulfill this demands while guaranteeing the maximum revenue. In this paper, we proposed a holistic view in how to achieve these two objectives through a formulation of the problem taking in the account the need for DC expansion while reducing the total operational cost through servers consolidation.

REFERENCES

[1] M. Wardat *et al.*, "To build or not to build? addressing the expansion strategies of cloud providers," in *FiCloud*. IEEE, 2014, pp. 477–482.
[2] M. Al-Ayyoub, M. Wardat, Y. Jararweh, and A. A. Khreishah, "Optimizing expansion strategies for ultrascale cloud computing data centers," *Simulation Modelling Practice and Theory*, vol. 58, pp. 15–29, 2015.
[3] P. J. Kuehn and M. E. Mashaly, "Automatic energy efficiency management of data center resources by load-dependent server activation and sleep modes," *Ad Hoc Networks*, vol. 25, pp. 497–504, 2015.
[4] S. Mazumdar and M. Pranzo, "Power efficient server consolidation for cloud data center," *FGCS*, vol. 70, pp. 4–16, 2017.
[5] J. Mate *et al.*, "Robust multi-tenant server consolidation in the cloud for data analytics workloads," in *ICDCS*. IEEE, 2017, pp. 2111–2118.
[6] A. Qureshi *et al.*, "Cutting the electric bill for internet-scale systems," *SIGCOMM Computer Communication Review*, 2009.
[7] E. Pinheiro, R. Bianchini, E. V. Carrera, and T. Heath, "Load balancing and unbalancing for power and performance in cluster-based systems," in *COLP*, vol. 180, 2001, pp. 182–195.
[8] M. Cardosa, M. R. Korupolu, and A. Singh, "Shares and utilities based power consolidation in virtualized server environments," in *Integrated Network Management, 2009. IM'09. IFIP/IEEE International Symposium on*. IEEE, 2009, pp. 327–334.
[9] A. Verma, P. Ahuja, and A. Neogi, "pmapper: power and migration cost aware application placement in virtualized systems," in *Middleware*. Springer, 2008, pp. 243–264.
[10] A. Gandhi, M. Harchol-Balter, R. Das, and C. Lefurgy, "Optimal power allocation in server farms," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 1. ACM, 2009, pp. 157–168.
[11] D. Kusic *et al.*, "Power and performance management of virtualized computing environments via lookahead control," *Cluster computing*, vol. 12, no. 1, pp. 1–15, 2009.
[12] S. Kundu *et al.*, "Revenue driven resource allocation for virtualized data centers," in *ICAC*. IEEE, 2015, pp. 197–206.
[13] M. Shojafar, C. Canali, R. Lancellotti, and E. Baccarelli, "Minimizing computing-plus-communication energy consumptions in virtualized networked data centers," in *ISCC*. IEEE, 2016, pp. 1137–1144.
[14] P. Arroba, J. M. Moya, J. L. Ayala, and R. Buyya, "Dynamic voltage and frequency scaling-aware dynamic consolidation of virtual machines for energy efficient cloud data centers," *CCPE*, 2016.
[15] S. Li, M. Brocanelli, W. Zhang, and X. Wang, "Data center power control for frequency regulation," in *PES*. IEEE, 2013, pp. 1–5.
[16] Z. Liu *et al.*, "Data center demand response: Avoiding the coincident peak via workload shifting and local generation," *Performance Evaluation*, vol. 70, no. 10, pp. 770–791, 2013.
[17] H. Chen, C. Hankendi, M. C. Caramanis, and A. K. Coskun, "Dynamic server power capping for enabling data center participation in power markets," in *ICCAD*. IEEE, 2013, pp. 122–129.
[18] C. Wang *et al.*, "A hierarchical demand response framework for data center power cost optimization under real-world electricity pricing," in *MASCOTS*. IEEE, 2014, pp. 305–314.
[19] J. Li *et al.*, "Towards optimal electric demand management for internet data centers," *IEEE TSG*, vol. 3, no. 1, pp. 183–192, 2012.
[20] X. Fan *et al.*, "Power provisioning for a warehouse-sized computer," in *SIGARCH Computer Architecture News*, 2007.
[21] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Energy-information transmission tradeoff in green cloud computing," *Carbon*, 2010.
[22] M. Ghamkhari and H. Mohsenian-Rad, "Energy and performance management of green data centers: A profit maximization approach," *IEEE TSG*, 2013.