A Game Theoretic Framework of SLA-Based Resource Allocation for Competitive Cloud Service Providers

Yanzhi Wang, Xue Lin, and Massoud Pedram
Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
{yanzhiwa, xuelin, pedram}@usc.edu

***Abstract*—Cloud computing is an emerging paradigm that allows the on-demand delivering of software, hardware, and data as services. It has attracted a lot of attention recently due to the increasing demand for high performance computing and storage. Resource allocation is one of the most important challenges in the cloud computing system, especially when the clients have some Service Level Agreements (SLAs) and the total profit depends on how the system can meet these SLAs. A set of multiple cloud service providers (CSPs) in the cloud, such as Google or Amazon, may support the similar type of application, and therefore, service requests generated from the network edges are free to be dispatched to any CSP in the set. This paper considers the problem of SLA-based resource provisioning and management among different CSPs. Each CSP owns a set of potentially heterogeneous servers supporting a common application type, and each performs resource allocation in these servers for request processing. In the cloud, a central request dispatcher allocates service requests to different servers (belonging to potentially different CSPs) based on the amounts of allocated resources in those servers. Each CSP optimizes its own profit, which is the total revenue obtained from servicing the clients subtracted by the total energy cost. The total revenue depends on the average service request response time as specified in the SLAs. The resource allocation problem among multiple CSPs forms a competitive normal-form game, since the payoff (profit) of each CSP depends not only on its own resource allocation results but also on the actions of the other CSPs. The existence and uniqueness of Nash equilibrium in this game are proved. Each CSP will find its optimal strategy at the Nash equilibrium point using the convex optimization technique. Experimental results demonstrate the effectiveness of the game theoretic resource provisioning framework for the CSPs.**

Keywords—cloud computing; cloud service provider; request dispatching; resource allocation; game theory

I. INTRODUCTION

Cloud computing has been widely envisioned as the next-generation computing paradigm for its advantages in location independent resource pooling, ubiquitous network access, on-demand service, and transference of risk [1]-[4]. Cloud computing shifts the computation and storage resources from the network edges to a "Cloud", from which businesses and users are able to access applications on demand from anywhere in the world. In cloud computing, the capabilities of various business applications are exposed as sophisticated services that can be accessed by the clients over a network. Cloud service providers (CSPs) are incentivized by the profits obtained from charging clients for accessing their services. Clients, on the other hand, are incentivized by the opportunity for enhancing performance and reducing the costs associated with "in-house" provisioning of these services. It is essential that the clients have guarantees from CSPs on the quality-of-service (QoS). Typically, the QoS requirements are specified in the Service Level Agreements (SLAs) brokered between CSPs and clients. The SLAs include requirements and guarantees on computing power, storage space, network bandwidth, availability and security, etc.

The underlying infrastructure of cloud computing is comprised of data centers and server clusters that are monitored and maintained by the CSPs [5]. The CSPs often end up over-provisioning their resources in these servers in order to meet the QoS requirements specified in SLAs [6]. Such over-provisioning will increase both the electrical energy cost and the carbon footprint incurred on the servers. Therefore, it is critical to perform optimal (*service*) *request dispatching* to various servers as well as optimal *resource allocation* in those servers in order to reduce the energy cost and the environmental impact, as have been investigated in [8] - [11]. The more general problem of resource allocation and management in distributed computing systems has been an active research topic in the past decade, in the context of grid computing systems [12][13], electronic commerce systems [14], autonomic computing systems [15][16], and in clusters of hosting servers [17][18].

Multiple CSPs in the cloud computing framework, such as Google and Amazon, may provide the same or similar type of applications as services e.g., web applications, large-scale scientific and engineering applications. For remote processing in the cloud, service requests can be dispatched to the servers of any CSP supporting such type of application. Each CSP in this framework will perform optimal resource provisioning in order for profit maximization. Game theoretic approaches are important to obtain a thorough analytical understanding of the resource provisioning problem among various CSPs. The authors [19] of considered the scenario where multiple CSPs cooperate with each other to establish a resource pool to support internal users and to offer services to public cloud users, and proposed a hierarchical cooperative game model and corresponding solutions. Similarly, reference [20] shows that

multiple CSPs can collaborate to establish a cloud federation, which in turn enhances the CSPs' ability to serve public cloud users. However, since most of the CSPs are entities aiming at profit maximization, it will be more realistic to assume that the CSPs in the cloud computing framework are non-cooperative (i.e., competitive) among each other.

In competitive games, one of the most widely utilized "solution concept" is the Nash equilibrium [30]. A set of strategies for the players constitute a Nash equilibrium if no player can benefit by changing his/her strategy unilaterally while the other players keep their strategies unchanged. In other words, every player is playing a *best response* to the strategy choices of his/her opponents. In the multiple-CSP framework, references [21][22] analyze a cloud computing system where different CSPs host their applications at an Infrastructure as a Service (IaaS) provider. The CSPs compete and bid for the usage of infrastructural resources in order to maximize their revenues from SLAs while minimizing the cost of resource usage. The service provisioning problem is modeled as a Generalized Nash game in [21][22], and run-time resource management policy is proposed for the CSPs.

In this work, we consider the problem of SLA-based resource provisioning and management among different CSPs in a cloud computing system. Different from references [21][22], each CSP owns a set of potentially heterogeneous servers (in terms of request processing capability) supporting a common type of application, and performs resource allocation in its servers for request processing. In the cloud, service requests from a common *request pool* are free to be dispatched to any server. A central request dispatcher allocates service requests to different servers (belonging to potentially different CSPs) based on the amounts of allocated resources in those servers. The total profit of each CSP for maximization is the total revenue obtained from request servicing, which depends on the average request response time as specified in the SLA, subtracted by the energy cost of the servers. We show that the resource allocation problem among multiple CSPs forms a competitive *normal-form* game, since the payoff (profit) of each CSP depends not only on its own resource allocation results but also on the actions of the other CSPs. We prove that this normal-form game is a strictly concave *n*-person game [31], and subsequently, prove the existence and uniqueness of the Nash equilibrium in this game. Each CSP will find its optimal strategy in the Nash equilibrium point using the convex optimization technique [28]. Experimental results demonstrate the effectiveness of the game theoretic resource provisioning optimization framework for the CSPs.

The rest of this paper is organized as follows. Section II introduces the system model for the game theory-based resource management problem in the cloud computing system. The game theoretic optimization problem formulation and optimization are provided in Section III. Experimental results are presented in Section IV, and we conclude this paper in Section V.

II. System Model and Problem Formulation

Figure 1 illustrates the structure of the target cloud computing resource allocation system, which is comprised of a service request pool, a central request dispatcher node, as well as a set of servers from N CSPs supporting the same type of application (e.g., web applications, large-scale scientific and engineering applications.) Each i^{th} CSP $(1 \le i \le N)$ owns and maintains M_i potentially heterogeneous servers. We use j as the index of servers of a CSP. Each j^{th} $(1 \le j \le M_i)$ server of the i^{th} CSP allocates a portion of its total resources, denoted by ϕ_{ij} $(0 \le \phi_{ij} \le 1)$, for servicing the requests. We use μ_{ij} to denote the average service request processing speed of the j^{th} server of the i^{th} CSP when all its resources have been allocated for request processing, i.e., $\phi_{ij} = 1$, and we name $\phi_{ij} \cdot \mu_{ij}$ the *computation resource* allocated by the corresponding server.

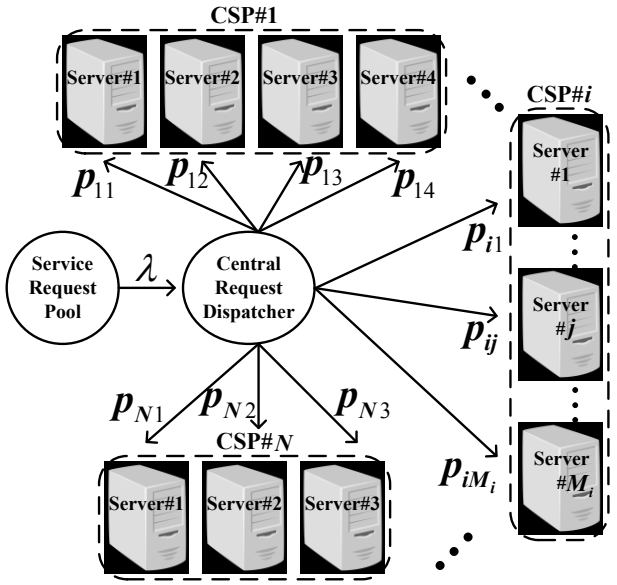

Fig. 1. Architecture of the Resource Allocation Problem in the Cloud Computing System with Multiple CSPs.

The service request pool contains service requests of a single type of application that are generated from all the clients. A service request is free to be dispatched to any server belonging to any CSP, because all the servers in the target cloud computing system can support such application type. As long as a service request is dispatched to a server, the server creates a dedicated virtual machine (VM) [23] for that service request, loads the application executable and starts execution. The central request dispatcher assigns a request to the the j^{th} server of the i^{th} CSP with probability p_{ij}, which is proportional to the amount of computation resource $\phi_{ij} \cdot \mu_{ij}$ allocated by that server. In other words, p_{ij} is given by:

$$p_{ij} = \frac{\phi_{ij} \cdot \mu_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \phi_{ij} \cdot \mu_{ij}} \quad (1)$$

In order the derive the analytical form of the average response time, service requests in the request pool are assumed to follow a Poisson process with an average generating rate of λ, which can be predicted based on the past behavior of the clients. According to the properties of the exponential distributions [26], service requests that are dispatched to the j^{th} server of the i^{th} CSP satisfies a Poisson process with an average rate of $p_{ij} \cdot \lambda$, which is the average request arrival rate of that server. Based on the well-known formula in the M/M/1 queues [27], the average response time of service requests dispatched to each j^{th} server of the i^{th} CSP is given by:

$$R_{ij} = \frac{1}{\phi_{ij} \cdot \mu_{ij} - p_{ij} \cdot \lambda} \tag{2}$$

Let P_{ij} denote the power consumption of each j^{th} server of the i^{th} CSP. P_{ij} is the sum of a constant term P_{ij}^{idle}, which represents the server power consumption when it is idle, and another term $P_{ij}^{act}(\phi_{ij})$ that is a superlinear function of the portion of allocated resources ϕ_{ij} for request processing. More specifically, P_{ij} is given by:

$$P_{ij} = P_{ij}^{idle} + P_{ij}^{act}(\phi_{ij}) \tag{3}$$

where the function $P_{ij}^{act}(\phi_{ij})$ can be presented as follows:

$$P_{ij}^{act}(\phi_{ij}) = f_{ij}^{act}(\phi_{ij}) \cdot (P_{ij}^{max} - P_{ij}^{idle}) \tag{4}$$

where P_{ij}^{max} is the server power consumption level when the server is active and all its resources have been allocated for service request processing i.e., $\phi_{ij} = 1$. $f_{ij}^{act}(\phi_{ij})$ is a normalized convex function of ϕ_{ij}, which equals to 0 when $\phi_{ij} = 0$ and equals to 1 when $\phi_{ij} = 1$. One typical function satisfying this property is $f_{ij}^{act}(\phi_{ij}) = (\phi_{ij})^2$. According to the power consumption expression in Eqn. (3), there is a most desirable utilization level ϕ_{ij} to optimize the P_{ij}/ϕ_{ij} value, which is unit power consumption level per allocated resource. The most desirable utilization level is typically around 70% in various references [24][25].

Let $U_i(R) = \beta_i - \gamma_i \cdot R$ denote the revenue the i^{th} CSP receives when servicing a request with response time equal to R, as specified in the SLA. Let $price_{ij}$ denote the unit electricity price at the location where the j^{th} server of the i^{th} CSP is built (i.e., we consider potential differences in electricity prices at different locations of the cloud computing framework.) Then the total profit of the i^{th} $(1 \le i \le N)$ CSP over a time period T is calculated by:

$$\begin{aligned} &\lambda \cdot T \cdot \sum_{j=1}^{M_i} p_{ij} \cdot \left(\beta_i - \gamma_i \cdot \frac{1}{\phi_{ij} \cdot \mu_{ij} - p_{ij} \cdot \lambda}\right) \\ &\quad - T \cdot \sum_{j=1}^{M_i} price_{ij} \cdot \left(P_{ij}^{idle} + P_{ij}^{act}(\phi_{ij})\right) \end{aligned} \tag{5}$$

where the first term of Eqn. (5) is the total revenue of the i^{th} CSP obtained from servicing the requests, as specified in the SLA, whereas the second term is the total energy cost over the time period T.

III. Optimiazation Problem Formulation and Solution

A. Problem Formulation

We consider the competition among the N CSPs in the cloud computing framework. Each CSP maximizes its own profit given in Eqn. (5). The optimization variables (action) of each i^{th} $(1 \le i \le N)$ CSP is the resource allocation vector $\boldsymbol{\phi}_i = \{\phi_{i1}, \phi_{i2}, \dots, \phi_{iM_i}\}$. As discussed in Section II, the request dispatching probability value p_{ij} depends not only on $\boldsymbol{\phi}_i$ of the i^{th} CSP but also on the actions of the other CSPs. Hence we denote the probability value as $p_{ij}(\boldsymbol{\phi}_i; \boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i}$ represents the resource allocation vectors of all the other CSPs than the i^{th} one. Of course, $\boldsymbol{\phi}_{-i}$ is not given to the i^{th} CSP when it makes decisions. The power consumption value P_{ij} in the j^{th} server of the i^{th} CSP, as shown in Eqn. (3), only depends on the local resource allocation action ϕ_{ij}. In summary, the payoff function (i.e., total profit) of the i^{th} $(1 \le i \le N)$ CSP to maximize, as given in Eqn. (5), can be represented as follows to emphasize the dependence on the optimization variables:

$$\begin{aligned} &\lambda \cdot T \cdot \\ &\sum_{j=1}^{M_i} p_{ij}(\boldsymbol{\phi}_i; \boldsymbol{\phi}_{-i}) \left(\beta_i - \gamma_i \cdot \frac{1}{\phi_{ij} \cdot \mu_{ij} - p_{ij}(\boldsymbol{\phi}_i; \boldsymbol{\phi}_{-i}) \cdot \lambda}\right) \\ &\quad - T \cdot \sum_{j=1}^{M_i} price_{ij} \cdot \left(P_{ij}^{idle} + P_{ij}^{act}(\phi_{ij})\right) \end{aligned} \tag{6}$$

Moreover, each i^{th} $(1 \le i \le N)$ CSP needs to satisfy the following constraints:

$$0 \le \phi_{ij} \le 1, \qquad \text{for } \forall j \tag{7}$$

$$\sum_{j=1}^{M_i} \phi_{ij} \cdot \mu_{ij} \ge \frac{\sum_{j=1}^{M_i} \mu_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \mu_{ij}} \cdot \lambda + \varepsilon \tag{8}$$

where ε is a small predefined value. Constraint (8) is enforced by the cloud computing framework to make sure that the all the requests from the service request pool can be serviced by certain CSP.

B. Optimization Procedure

Because the payoff function (6) of each i^{th} $(1 \le i \le N)$ CSP depends on not only its own action $\boldsymbol{\phi}_i$ but also the actions of the other CSPs, the resource management problem stated in Section III.A essentially forms a non-cooperative *normal-form* game, where all the players take action simultaneously. We name the normal-form game the *Resource Allocation among*

Multiple Cloud Service Providers (RA-MCSP) game. The players in the RA-MCSP game are the N CSPs in the cloud computing framework. The strategy of each i^{th} $(1 \le i \le N)$ CSP is the $\boldsymbol{\phi}_i$ vector, and the constraints are given in Eqns. (7), (8).

As the CSPs in the cloud computing framework are considered to be non-cooperative among each other, we are interested in the existence and uniqueness of the Nash equilibrium [30]. As one of the most widely utilized "solution concept" in normal-form games, the Nash equilibrium is the optimal strategy profile for all the players in the sense that no player can benefit by changing his/her strategy unilaterally while the other players keep their strategies unchanged. In other words, no player (CSP) will have incentive to leave the current strategy in the Nash equilibrium. We prove the existence and uniqueness of the Nash equilibrium in the RA-MCSP game.

Theorem I (Nash equilibrium in the RA-MCSP game): The Nash equilibrium in the RA-MCSP game exists and is unique.

Proof: We are going to prove that the RA-MCSP game is a strictly concave *n*-person game. We need to prove (i) the domain of the strategy profile for all the players, which is specified by constraints (7), (8), is a closed convex set, and (ii) the objective (payoff) function of each player to maximize is a strictly concave function with respect to the optimization variables of that player, assuming that the optimization variable values of the other players are given in prior. One can easily observe that statement (i) is true because constraints (7), (8) are all linear constraints of optimization variables $\boldsymbol{\phi}_i$. In the following, we prove that statement (ii) is also true:

- The first term of the payoff function (6) is a concave function of the optimization variables $\boldsymbol{\phi}_i$ as long as (i) $\sum_{j=1}^{M_i} p_{ij}(\boldsymbol{\phi}_i; \boldsymbol{\phi}_{-i})$ is a concave function of $\boldsymbol{\phi}_i$ when $\boldsymbol{\phi}_{-i}$ is given, and (ii) $\frac{p_{ij}(\boldsymbol{\phi}_i;\boldsymbol{\phi}_{-i})}{\phi_{ij}\cdot\mu_{ij}-p_{ij}(\boldsymbol{\phi}_i;\boldsymbol{\phi}_{-i})\cdot\lambda}$ is a convex function of $\boldsymbol{\phi}_i$. We prove statement (i) as follows:

$$\begin{aligned}\sum_{j=1}^{M_i} p_{ij}(\boldsymbol{\phi}_i; \boldsymbol{\phi}_{-i}) &= \sum_{j=1}^{M_i} \frac{\phi_{ij}\cdot\mu_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\phi_{ij}\cdot\mu_{ij}} \\ &= \frac{\sum_{j=1}^{M_i}\phi_{ij}\cdot\mu_{ij}}{\sum_{j=1}^{M_i}\phi_{ij}\cdot\mu_{ij}+\sum_{i'\neq i}\sum_{j=1}^{M_{i'}}\phi_{i'j}\cdot\mu_{i'j}} \\ &= 1-\frac{\sum_{i'\neq i}\sum_{j=1}^{M_{i'}}\phi_{i'j}\cdot\mu_{i'j}}{\sum_{j=1}^{M_i}\phi_{ij}\cdot\mu_{ij}+\sum_{i'\neq i}\sum_{j=1}^{M_{i'}}\phi_{i'j}\cdot\mu_{i'j}}\end{aligned} \tag{9}$$

  Hence we have proved the statement (i) because $\sum_{i'\neq i}\sum_{j=1}^{M_{i'}}\phi_{i'j}\cdot\mu_{i'j}$ is assumed to be a constant value. Moreover, we prove statement (ii) as follows:

$$\begin{aligned}&\frac{p_{ij}(\boldsymbol{\phi}_i;\boldsymbol{\phi}_{-i})}{\phi_{ij}\cdot\mu_{ij}-p_{ij}(\boldsymbol{\phi}_i;\boldsymbol{\phi}_{-i})\cdot\lambda} \\ &= \frac{\frac{\phi_{ij}\cdot\mu_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\phi_{ij}\cdot\mu_{ij}}}{\phi_{ij}\cdot\mu_{ij}-\frac{\phi_{ij}\cdot\mu_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\phi_{ij}\cdot\mu_{ij}}\cdot\lambda} \\ &= \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\phi_{ij}\cdot\mu_{ij}-\lambda} \\ &= \frac{1}{\sum_{j=1}^{M_i}\phi_{ij}\cdot\mu_{ij}+\sum_{i'\neq i}\sum_{j=1}^{M_{i'}}\phi_{i'j}\cdot\mu_{i'j}-\lambda}\end{aligned} \tag{10}$$

  And we have proved the convexity of the function $\frac{p_{ij}(\boldsymbol{\phi}_i;\boldsymbol{\phi}_{-i})}{\phi_{ij}\cdot\mu_{ij}-p_{ij}(\boldsymbol{\phi}_i;\boldsymbol{\phi}_{-i})\cdot\lambda}$.

- The second term of the payoff function (6), i.e., $T\cdot\sum_{j=1}^{M_i} price_{ij}\cdot\left(P_{ij}^{idle}+P_{ij}^{act}(\phi_{ij})\right)$, is a strictly convex function of the optimization variable $\boldsymbol{\phi}_i$, because each function $P_{ij}^{act}(\phi_{ij})$ is a strictly convex function of ϕ_{ij}.

After we have proved that the RA-MCSP game is a strictly concave *n*-person game, the existence and uniqueness of Nash equilibrium are directly resulted from the first and third theorems in [31], respectively. ■

Each i^{th} player (CSP) of the RA-MCSP game finds its optimal strategy in the Nash equilibrium using standard convex optimization technique [28]. The detailed procedure is illustrated in Algorithm 1.

Algorithm 1: Finding the Nash Equilibrium in the RA-MCSP Game for the i^{th} CSP.

Initialize the $\boldsymbol{\phi}_i$ vector (i.e., the resource allocation results of the i^{th} CSP), as well as $\boldsymbol{\phi}_{-i}$ for the other CSPs, satisfying constraints (7), (8).

Do the following procedure iteratively:

> **For each** $1 \le i' \le N$:
>
> > Find the optimal $\boldsymbol{\phi}_{i'}$ vector (i.e., the *best response* of CSP i') with respect to $\boldsymbol{\phi}_{-i'}$, by solving the convex optimization problem with objective function (6) and constraints (7), (8) using standard techniques [29].
> >
> > Update the $\boldsymbol{\phi}_{i'}$ vector to be the new value.
>
> **End**

Until the solution converges.

Return the optimized $\boldsymbol{\phi}_i$ vector.

IV. EXPERIMENTAL RESULTS

In this section, we implement the game theory-based resource provisioning framework for multiple CSPs, and compare the optimization results with baseline resource allocation algorithms. We use normalized amounts of most of the parameters in the cloud computing system instead of their real values.

We consider a cloud computing framework that is comprised of five CSPs. The five CSPs contain 4 servers, 6 servers, 5 servers, 7 servers, and 3 servers, respectively. The average service request generating rate λ is the parameter that we sweep in the experiments. The average service request processing rate μ_{ij} in each j^{th} $(1 \le j \le M_i)$ server of the i^{th} $(1 \le i \le N)$ CSP is a uniformly distributed random variable between 8 and 12. The maximum power consumption P_{ij}^{max} of each server is a uniformly distributed random variable between 250 and 350. The idle power consumption P_{ij}^{idle} of each server is uniformly distributed between 60 and 20. The normalized superlinear function $f_{ij}^{act}(\phi_{ij})$ is given by $f_{ij}^{act}(\phi_{ij}) = (\phi_{ij})^2$. The unit energy price value $price_{ij}$ is uniformly distributed between 0.1 and 0.2, at the building location of each server in the cloud. For the utility functions of each i^{th} $(1 \le i \le N)$ CSP, parameters β_i and γ_i are uniformly distributed between 10 and 12, and between 4 and 6, respectively. The time period T is assumed to be 1, i.e., the unit time period.

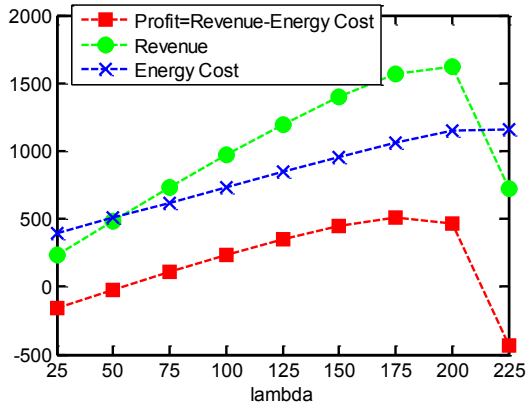

Fig. 2. The Total Revenue, Total Energy Cost, and Total Profit of the Cloud Computing System with Different λ Values.

In the first experiment, we change the average service request generating rate λ from 25 to 225, and observe the total revenue, total energy cost, and total profit (revenue – energy cost) of all the five CSPs, as illustrated in Figure 2. In this experiment, each CSP is a rational player and finds its best strategy in the Nash equilibrium of the RA-MCSP game by executing Algorithm 1. We can observe from Figure 2 that (i) the total energy cost of the cloud computing system keeps increasing when λ increases because of the increasing amount of resource required for service request processing, and (ii) the total revenue from servicing requests and the total profit of the cloud computing system first increase when λ increases due to the increasing number of served requests, and then decrease during the further increasing of the parameter λ. This phenomenon is because of the increasing of the request response time significantly reduces the revenue of each CSP (to be even negative) as specified in the SLA. As shown in Figure 2, the overall cloud computing system achieves the maximum profit when $\lambda = 175$.

In the second experiment, we compare the profit of the 3rd CSP (with 5 servers) in this framework achieved by executing the proposed game theoretic optimization method (Algorithm 1) and three baseline methods. In this experiment, all the other CSPs are rational and find the best strategy in the Nash equilibrium of the RA-MCSP game by executing Algorithm 1. In the three baseline systems, the 3rd CSP is not aware of the optimal strategy in the Nash equilibrium. Instead, it uses estimation of the other CSPs' strategies (resource allocation results). In Baseline 1, the 3rd CSP assumes that all the other CSPs allocate all the server resources for request processing (i.e., the ϕ_{ij} values are equal to 1 for $i = 1,2,4,5$), and performs optimal resource allocation (i.e., finding the optimal ϕ_{3j} values) based on this assumption. In Baseline 2 and 3, the 3rd CSP calculates the minimum resource required in each j^{th} server of the i^{th} CSP as follows:

$$\phi_{min} = \frac{\lambda}{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\mu_{ij}} \tag{11}$$

One can observe that the underlying assumption is to allocate the same portion of resource in each server of the cloud. Then Baseline 2 assumes that all the other CSPs allocate $1.2\phi_{min}$ portion of resource in their servers, whereas Baseline 3 assumes $1.5\phi_{min}$ portion of resource for request processing. Of course the portion of resource allocation cannot exceed 100%. Then the 3rd CSP of our interest performs optimal resource allocation based on the corresponding assumptions.

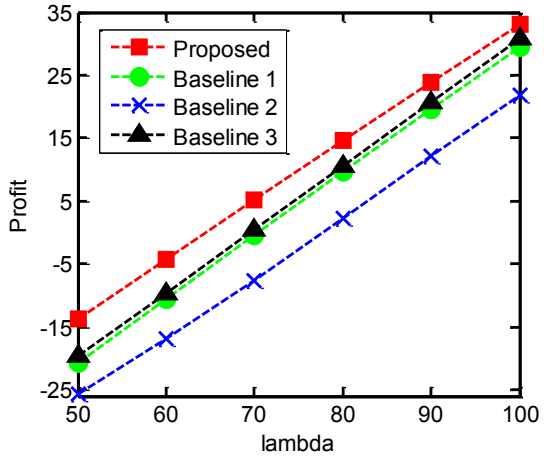

Fig. 3. The Profit of the 3rd CSP Achieved by the Proposed Method and Three Baseline Algorithms when the λ Value Increases from 50 to 100.

Figure 3 illustrates the profit of the 3rd CSP achieved by the proposed method and three baseline algorithms when the λ value increases from 50 to 100. We can observe that (i) the game theoretic optimization method consistently results in higher profit for the 3rd CSP compared with baseline methods, illustrating that the CSP finds its best response using the Nash equilibrium-based optimization method, and (ii) the profit gain achieved by the proposed method gradually reduces with the increase of λ. For example, the 3rd CSP achieves 13.5X profit when $\lambda = 70$ by employing the game theoretic optimization method compared with Baseline 3 (the best-performing baseline method.) This profit gain reduces to 36.8% when $\lambda = 80$ and to 14.9% when $\lambda = 90$.

V. CONCLUSION

In this paper, we consider the problem of SLA-based resource provisioning problem among different CSPs in the cloud computing framework. Each CSP hosts a set of potentially heterogeneous servers and performs resource allocation in these servers for request processing. In the cloud, service requests from a common request pool are free to be dispatched to any server. A central request dispatcher allocates service requests to different servers (belonging to potentially different CSPs) based on the amounts of allocated resources in those servers. The objective of each CSP is to maximize its own profit, which is the total revenue obtained from request servicing subtracted by the energy cost of the servers. The total revenue depends on the average service request response time as specified in the SLAs. We show that the resource provisioning problem among multiple CSPs forms a competitive normal-form game, since the payoff (profit) of each CSP depends not only on its own resource allocation results but also on the actions of the other CSPs. We prove that this normal-form game is a strictly concave n-person game, and subsequently, prove the existence and uniqueness of the Nash equilibrium in this game. Each CSP will find its optimal strategy in the Nash equilibrium point using the convex optimization technique. Experimental results demonstrate the effectiveness of the game theory-based resource provisioning optimization framework for the CSPs.

REFERENCES

[1] B. Hayes, "Cloud Computing," *Communications of the ACM*, 2008.

[2] R. Buyya, "Market-oriented cloud computing: vision, hype, and reality of delivering computing as the 5th utility, " in *IEEE/ACM International Symposium on Cluster Computing and the Grid* (CCGrid), 2009.

[3] M. Pedram, "Energy-efficient datacenters," *IEEE Trans. on CAD*, 2012.

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Pabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Communications of the ACM*, 2010.

[5] R. H. Katz, "Tech Times Building Boon," *IEEE Spectrum*, 2009.

[6] L. A. Barroso and U. Holzle, "The case for energy-proportional computing," *IEEE Computer*, 2007.

[7] H. Goudarzi and M. Pedram, "Multi-dimensional SLA-based resource allocation for multi-tier cloud computing systems," *Proc. of IEEE Cloud Computing Conference* (CLOUD), 2011.

[8] Y. Wang, S. Chen, H. Goudarzi, and M. Pedram, "Resource allocation and consolidation in a multi-core server cluster using a Markov decision process model," *Proc. of Internatinoal Symposium on Quality Electronic Design* (ISQED), 2013.

[9] Y. Wang, X. Lin, and M. Pedram, "A sequential game perspective and optimization of the smart grid with distributed data centers," *Proc. of IEEE PES Innovative Smart Grid Technologies Conference* (ISGT), 2013.

[10] Y. Wang, S. Chen, and M. Pedram, "Service level agreement-based joint application environment assignment and resource allocation in cloud computing systems," in *IEEE Green Technologies Conference* (GreenTech), 2013.

[11] G. Wei, A. V. Vasilakos, Y. Zheng, and N. Xiong, "A game-theoretic method for fair resource allocation for cloud computing services," *The Journal of Supercomputing*, 2010.

[12] K. Kraute, R. Buyya, and M. Maheswaran, "A taxonomy and survey of grid resource management systems for distributed computing," *Software Practice and Experience*, 2002.

[13] R. Buyya and M. Murshed, "GridSim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing," *Concurrency and Computation Practice & Experience*, 2002.

[14] Z. Liu, M. S. Squillante, and J. L. Wolf, "On maximizing service-level agreement profits," in *3rd ACM Conference on Electronic Commerce*, 2001.

[15] D. Ardagna, M. Trubian, and L. Zhang, "SLA based resource allocation policies in autonomic environments," *Journal of Parallel and Distributed Computing*, 2007.

[16] L. Zhang and D. Ardagna, "SLA based profit optimization in autonomic computing systems," in *2nd Int. Conf. on Service Oriented Computing*, 2004.

[17] A. Chandra, W. Gong, and P. Shenoy, "Dynamic resource allocation for shared clusters using online measurements," *International Conference on Measurement and Modeling of Computer Systems* (SIGMETRICS), 2003.

[18] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, and R. P. Doyle, "Managing energy and server resources in hosting centers," in *Proc. of the 18th ACM Symposium on Operating Systems Principles* (SOSP'01), 2001.

[19] D. Niyato, A. V. Vasilakos, and K. Zhu, "Resource and revenue sharing with coalition formulation of cloud providers: game theoretic approach," in *Proc. of IEEE/ACM International Symposium on Cluster Computing and the Grid* (CCGrid), 2011.

[20] Z. Lu, X. Wen, and Y. Sun, "A game theory based resource sharing scheme in cloud computing environment," in *Proc. of World Congress on Information and Communication Technologies* (WICT), 2012.

[21] D. Ardagna, B. Panicucci, and M. Passacantando, "A game theoretic formulation of the service provisioning problem in cloud systems," in *Proc. of the 20th International Conference on World Wide Web*, 2011.

[22] D. Ardagna, B. Panicucci, and M. Passacantando, "Generalized Nash equilibria for the service provisioning problem in cloud systems," *IEEE Trans. on Services Computing*, 2012.

[23] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the art of virtualization," in *Proc. of the 19th ACM Symposium on Operating System Principles* (SOSP), 2003.

[24] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," in *Proc. of Conf. on Power Aware Computing and Systems* (HotPower'08), 2008.

[25] Y. Gao, Y. Wang, S. K. Gupta, and M. Pedram, "An energy and deadline aware scheduling and optimization framework for cloud service providers," in *Proc. of the International Conference on Hardware/Software Codesign and System Synthesis* (CODES+ISSS), 2013.

[26] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 3rd edition, 1991.

[27] L. Kleinrock, *Queueing Systems, Volume I: Theory*, New York: Wiley, 1975.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[29] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 1.21.” http://cvxr.com/cvx, Feb. 2011.

[30] K. Leyton-Brown and Y. Shoham, *Essentials of Game Theory: A Concise, Multidisciplinary Introduction, Morgan & Claypool Publishers*, 2008.

[31] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave *n*-person games," *Econometrica*, vol. 33, pp. 347 - 351, 1965.