Optimal Admission Control Policy for Mobile Cloud Computing Hotspot with Cloudlet

Dinh Thai Hoang, Dusit Niyato, and Ping Wang
School of Computer Engineering, Nanyang Technological University (NTU), Singapore

***Abstract*—We consider an admission control problem and adaptive resource allocation for running mobile applications on a cloudlet. We formulate an optimization problem for dynamic resource sharing of mobile users in mobile cloud computing (MCC) hotspot with a cloudlet as a semi-Markov decision process (SMDP). SMDP is transformed into a linear programming (LP) model and it is solved to obtain an optimal solution. In the optimization model, the quality of service (QoS) for different classes of mobile user is taken into account under resource constraints (i.e., bandwidth and server). The numerical results are presented to illustrate that the proposed admission control scheme can achieve a desirable performance and improve throughput of an MCC hotspot significantly.**

Keywords-Mobile Cloud, Cloudlet, admission control, Markov Decision Process, blocking probability, resource allocation.

I. Introduction

Mobile cloud computing (MCC) can provide optimal services for customers through taking advantages of both mobile service and cloud computing. In MCC, data processing and storage for mobile users will be provided as services on the cloud. As a result, mobile device do not need a powerful configuration since all the complicated computing modules, which usually require high-speed CPU and large memory size, and consume a lot of energy from battery, can be processed in the cloud. Many applications have been developed based on MCC (e.g., healthcare [1], education [2], and multimedia [3]). However, MCC faces long latency problem because of connection to remote servers in the cloud [4]. Therefore, *cloudlet* has been proposed as a solution for this problem.

Cloudlet is a trusted, resource-rich computer of computers which is well-connected to the Internet and available for use by nearby mobile devices (e.g., in coffeeshop or library). Hence, when mobile devices cannot or do not want to connect to the cloud, they can find and access a nearby computing resource. In this way, mobile users may be met the demand for real-time interactive response by low-latency, one-hop, high-bandwidth wireless access of a hotspot to the cloudlet [5]. However, a cloudlet does not have abundant resource as that of the cloud. Therefore, how to allocate resources efficiently is a concern in this MCC model.

There are some works considering a resource allocation problem of MCC in the literature. A hierarchical cloud computing architecture was introduced and a resource management mechanism for cloudlets was proposed in [6]. In [7], the cloudlet was used to support CPU and GPU (Graphics Processing Unit) sharing among users running a graphic application. The design of cloudlet was introduced to support highly energy-efficient, secure, reliable, and low total-ownership cost of the computing infrastructure. [8] considered a resource allocation problem for the security services in cloud computing. However, the proposed scheme limits the users to be in two classes only and it also does not consider priority of users in different classes that is very important in real systems. In practice, users have different demands in using resource on the cloudlet and all their works on the cloudlet will be charged by the service provider. So that, if we do not classify users, it is very hard for service providers to charge customers and the system also cannot provide high quality of service (QoS) for them. The bandwidth capacity which was not taken into account is an important issue for resource-limited mobile device used in MCC. To the best of our knowledge, the development of optimal admission control schemes for cloudlet using rigorous mathematical modeling has not been well investigated in the literature.

In this paper, we develop an optimization model to address the admission control issue for the MCC hotspot with a cloudlet. We propose an optimization model based on the semi-Markov decision process (SMDP) [9] to maximize the reward (e.g., revenue of service provider) of the resource usage in MCC hotspot while still meeting the QoS requirements of mobile users. In particular, we consider SMDP with multi-constraints including constraints about resource, bandwidth, and QoS. The policy of SMDP can be obtained by transforming an original optimization problem into a linear programming (LP) model in which the resource and QoS constraints (e.g., bandwidth and blocking probability) can be included. The LP model can be solved efficiently by using standard solver. The use of LP framework is very appropriate for the simple system (i.e., the state space and action space are not huge) since it can produce the exact optimal policy and be easy to add more constraints for the system. In addition, this framework can be calculated offline and be applied for online admission control of MCC hotspot. The numerical results clearly show that the QoS performance of users in different class is ensured. Also, the reward and resource utilization are maximized.

The remaining of this paper is organized as follows. In Section II, we present the system model and state different constraints of the admission control problem. Optimal solution for this problem is developed by using SMDP theory in Section III. We show the numerical results in Section IV and summary the paper in Section V.

978-1-4673-0437-5/12/$31.00 ©2012 IEEE

II. System Model

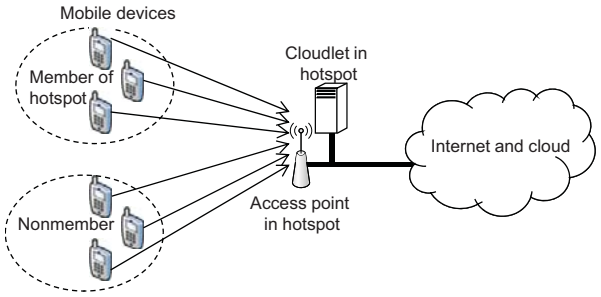

Fig. 1. System model of mobile cloud computing hotspot.

We consider a mobile cloud computing (MCC) service in a hotspot as shown in Fig. 1. The MCC hotspot provides a wireless access and it also has a cloudlet to serve its customers running mobile applications. In this case, mobile users connect to the access point and offload the computing modules of the mobile application to a cloudlet. There are C classes of mobile users whose set is denoted by $\{1, \ldots, c, \ldots, C\}$. Each class of users has a certain priority level. For example, in Fig. 1, there are two classes of users, namely "member" and "nonmember". Members are the users who subscribe to the MCC service with a hotspot in advance, and hence they have higher priority of using resource (i.e., bandwidth and resources on server) of MCC hotspot. On the other hand, nonmembers are the users who request the MCC services from the hotspot in an on-demand fashion, and they have lower priority. In this case, the different priority is differentiated by QoS performance in terms of blocking probability. For instance, member users are expected to experience smaller blocking probability than that of nonmembers. There are S types of services whose set is denoted by $\{1, \ldots, s, \ldots, S\}$. Different type of service corresponds to different mobile application which requires different amount of resource.

There are many applications which can be deployed in MCC as discussed in [4]. Because of limited resource of mobile devices, mobile users want to offload these applications to the cloud to save battery lifetime and to improve the efficiency as well as running speed. However, offloading is not always the effective way to save energy [10]. Thus, before offloading for mobile applications, the mobile users need to determine whether to offload and which portions of the application should be offloaded. After partitioning the application and determining which partitions need to be offloaded, the mobile device will send these partitions to the cloudlet for processing. Each partition performed on the cloudlet is considered as a service provided by the cloudlet and this application partition occupies a certain amount of resource of the cloudlet when it is executed. Therefore, we can assume that the total resource of a cloudlet is denoted by N (e.g., CPU share, memory, and disk capacity). Let n_s denote the cloudlet resource used by one mobile user for service s. Let x_c^s denote the number of mobile users in class c for service s in MCC hotspot. This number of mobile users represents the state of the MCC hotspot. The resource constraint for a cloudlet can be expressed as follows:

$$N \geq \sum_{s=1}^{S} \sum_{c=1}^{C} n_s x_c^s. \tag{1}$$

This constraint guarantees that the resource used by mobile users does not exceed the capability of the cloudlet. For an access point which provides a wireless connection for mobile users in a hotspot, the total bandwidth is denoted by B. Let b_s denote the bandwidth required by one mobile user for service s. The resource constraint for an access point is expressed as follows:

$$B \geq \sum_{s=1}^{S} \sum_{c=1}^{C} b_s x_c^s. \tag{2}$$

This constraint ensures that the bandwidth used by mobile users does not exceed the capacity of the access point.

The arrival process of mobile users in class c for service s is assumed to follow the Poisson distribution with mean rate λ_c^s and the cloudlet resource occupation time follows the exponential distribution with mean $1/\mu_c^s$. An admission control mechanism of the MCC hotspot is designed to decide whether an arriving request from mobile user can be accepted or not. The admission control mechanism takes the state of MCC hotspot into account and makes decision according to the optimal policy to be obtained from SMDP. The detail of SMDP formulation is provided in the next section.

III. Optimal Resource Allocation Policy

In this section, we formulate the admission control problem of an MCC hotspot as a semi-Markov decision process (SMDP) [9]. In the following, we will define the state space, action space and reward function for the underlying SMDP. We will determine an optimal admission control policy as the optimal solution of this SMDP under some given constraints through using linear programming (LP) approach.

A. Description of SMDP

Decision epoch of the underlying SMDP is arrival and departure instant of mobile user in class c for service s. The current state at decision epoch of the system (i.e., MCC hotspot) will be represented by a vector $\mathbf{x}$ as follows:

$$\mathbf{x} \triangleq \left[x_1^1, \ldots, x_1^S, \ldots, x_C^1, \ldots, x_C^S\right] \tag{3}$$

where x_c^s is the number of ongoing users in class c for service s in the MCC hotspot. Then, the state space $\mathcal{X}$ is defined as follows:

$$\mathcal{X} \triangleq \left\{ \mathbf{x} : \mathbf{x} \geq 0; N \geq \sum_{s=1}^{S} \sum_{c=1}^{C} n_s x_c^s, B \geq \sum_{s=1}^{S} \sum_{c=1}^{C} b_s x_c^s \right\}. \tag{4}$$

In particular, the state space is constrained by the resource and bandwidth of a cloudlet and access point as defined in (1) and (2), respectively. The variable e defined as follows:

$$e \in \{a_1^1, \ldots, a_1^S, \ldots, a_C^1, \ldots, a_C^S\}$$

represents the event type of a request arrival. The indicator a_c^s is the admission control action when a request arrival happens from mobile user in class c for service s in the MCC hotspot. The admission control action is defined as follows:

$$a_c^s = \begin{cases} 1, & \text{if an arrival request for service } s \text{ is accepted} \\ 0, & \text{otherwise.} \end{cases}$$

In this case, when the MCC hotspot is in state $\mathbf{x}$, an accept/reject decision of admission control of MCC hotspot must be made for each type of possible arrival. Hence, the action space can be defined as follows:

$$\mathcal{A} \triangleq \{(a_1^1, \ldots, a_1^S, \ldots, a_C^1, \ldots, a_C^S) : a_c^s \in \{0,1\}\} \quad (5)$$

$$\text{for } c = 1, 2, \ldots, C; s = 1, 2, \ldots, S.$$

Moreover, admissible action space $\mathcal{A}_{\mathbf{x}}$ given a system state $\mathbf{x}$ comprises all possible actions that do not result in transition into a state that is not allowed (i.e., the state must be in the space $\mathcal{X}$ as defined in (4)). In addition, for state $\mathbf{x}_0 = [0, \ldots, 0]$, it is required that action $a = 0$ is excluded from $\mathcal{A}_{\mathbf{x}_0}$ to prevent the system to be trapped in this state $\mathbf{x}_0$ forever. Therefore, the action space is actually a state-dependent subset defined by

$$\mathcal{A}_{\mathbf{x}} \triangleq \{a \in \mathcal{A} : a_c^s = 0 \text{ if } \mathbf{x} + e_c^s \notin \mathcal{X}\} \quad (6)$$

where e_c^s is a vector having all zeros except the one at the same position as x_c^s in (3).

We now analyze the dynamics of this SMDP, which is characterized by the state transition probabilities of the Markov chain obtained by embedding the system at arrival and departure instant. Specifically, we will determine transition probability $p_{\mathbf{xy}}(a)$ from state $\mathbf{x}$ to state $\mathbf{y}$ when action a is taken. Let $\tau_{\mathbf{x}}(a)$ denote the expected time until the next decision epoch after action a is taken at state $\mathbf{x}$. Then, $\tau_{\mathbf{x}}(a)$ can be calculated as the inverse of the cumulative arrival and departure rate with blocked arrivals taken into account. In particular, $\tau_{\mathbf{x}}(a)$ can be calculated as follows:

$$\tau_{\mathbf{x}}(a) = \left[\sum_{c=1}^{C} \sum_{s=1}^{S} \lambda_c^s a_c^s + \sum_{c=1}^{C} \sum_{s=1}^{S} \mu_c^s x_c^s \right]^{-1} \quad (7)$$

for $a \in \mathcal{A}$. We are now ready to calculate transition probability $p_{\mathbf{xy}}(a)$ of the underlying embedded Markov chain. This can be done by noting that the probability of a certain event (e.g., connection arrival and departure) is equal to the ratio between the rate of that event and the total cumulative event rate $1/\tau_{\mathbf{x}}(a)$. Then, the transition probability $p_{\mathbf{xy}}(a)$ can be expressed as follows:

$$p_{\mathbf{xy}}(a) = \begin{cases} \lambda_c^s a_c^s \tau_{\mathbf{x}}(a), & \text{if } \mathbf{y} = \mathbf{x} + e_c^s \\ \mu_c^s x_c^s \tau_{\mathbf{x}}(a), & \text{if } \mathbf{y} = \mathbf{x} - e_c^s \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Let $r(\mathbf{x}, a)$ be the reward rate when the MCC hotspot is in state $\mathbf{x}$ and action a has been chosen. If r_c^s is the reward rate of mobile user in class c for service s, then the total reward for the MCC hotspot is calculated by

$$r(\mathbf{x}, a) = \sum_{c=1}^{C} \sum_{s=1}^{S} r_c^s x_c^s. \quad (9)$$

In the following, we formulate the optimal admission control problem using the above description of the underlying SMDP.

B. Formulation of Admission Control Problem

The MCC hotspot optimizes the resource allocation policy π for mobile users (i.e., whether to accept an arrival request from a mobile user or not). This policy π is a mapping from state $\mathbf{x} \in \mathcal{X}$ to action $a \in \mathcal{A}$ to maximize the average reward of MCC hotspot, i.e.,

$$\max_{z_{\mathbf{x},a} \geq 0} \quad \sum_{\mathbf{x} \in \mathcal{X}} \sum_{a \in \mathcal{A}_{\mathbf{x}}} r(\mathbf{x}, a) \tau_{\mathbf{x}}(a) z_{\mathbf{x},a} \quad (10)$$

subject to the constraints:

$$\sum_{\mathbf{x} \in \mathcal{X}} \sum_{a \in \mathcal{A}_{\mathbf{x}}} z_{\mathbf{x},a} \tau_{\mathbf{x}}(a) = 1$$

$$\sum_{a \in \mathcal{A}_{\mathbf{y}}} z_{\mathbf{y},a} - \sum_{\mathbf{x} \in \mathcal{X}} \sum_{a \in \mathcal{A}_{\mathbf{x}}} p_{\mathbf{xy}}(a) z_{\mathbf{x},a} = 0, \mathbf{y} \in \mathcal{X}$$

$$z_{\mathbf{x},a} \geq 0, \mathbf{x} \in \mathcal{X}, a \in \mathcal{A}_{\mathbf{x}}. \quad (11)$$

In (11), the term $z_{\mathbf{x},a}\tau_{\mathbf{x}}(a)$ represents the steady-state probability at which the system is in state $\mathbf{x}$ and action a is taken. Hence, the first constraint in (11) requires that sum of the steady-state probabilities should be equal to 1 and the second constraint can be interpreted as a balance equation.

By solving the linear programming (LP) formulation defined in (10)-(11), we can obtain an optimal admission control policy. However, we also need to consider the QoS requirements, i.e., the upper-bound for the blocking probability. Let P_b^c denote the maximum tolerable blocking probability of mobile users in class c. Then, the following constraint needs to be added into LP formulation defined in (10)-(11), i.e.,

$$\sum_{\mathbf{x} \in \mathcal{X}} \sum_{a \in A} \sum_{s=1}^{S} (1 - a_c^s) z_{\mathbf{x},a} \tau_{\mathbf{x}}(a) \leq P_b^c$$

$$\text{for } c = 1, 2, \ldots, C. \quad (12)$$

In addition, we also need to consider the available bandwidth from an access point in an MCC hotspot. Therefore, to make a stable system, the total bandwidth usage of an MCC hotspot has to be lower than or equal to the bandwidth of the corresponding access point. Recall that B is the bandwidth of access point and b_s is the required bandwidth for a mobile user for service s. The constraint of bandwidth can be expressed as follows:

$$\sum_{\mathbf{x} \in \mathcal{X}} \sum_{a \in A} \sum_{s=1}^{S} b_s z_{\mathbf{x},a} \tau_{\mathbf{x}}(a) \leq B. \quad (13)$$

The optimal admission control policy can be obtained as follows. We calculate optimal $z_{\mathbf{x},a}^*$ by solving the LP formulation defined in (10)-(11) along with the constraints defined

in (12) and (13). Then, we can obtain an optimal randomized admission control policy. The optimal randomized admission control policy for each state $\mathbf{x}$ is the probability of choosing action $a \in \mathcal{A}_{\mathbf{x}}$. This probability can be calculated as follows:

$$\theta_{\mathbf{x}}(a) = z^*_{\mathbf{x},a} / \sum_{a' \in \mathcal{A}_{\mathbf{x}}} z^*_{\mathbf{x},a'}. \quad (14)$$

Note that the optimal randomized admission control policy can be obtained offline and applied for online admission control of MCC hotspot given the different states.

Various performance measures can be obtained as follows:

- *Blocking probability:* Blocking probability determines the chance that a request from a mobile user to access MCC hotspot will be rejected. This blocking probability can be obtained as follows:

$$P_c = \sum_{\mathbf{x} \in \mathcal{X}} \sum_{a \in A} \sum_{s=1}^{S} (1 - a_c^s) z^*_{\mathbf{x},a} \tau_{\mathbf{x}}(a). \quad (15)$$

- *Throughput:* The throughput of the MCC hotspot is the number of mobile users who are successfully accepted into MCC hotspot per unit of time, and it can be obtained from

$$\overline{T} = \sum_{c=1}^{C} \sum_{s=1}^{S} (1 - P_c) \lambda_c^s. \quad (16)$$

- *Average bandwidth usage:* The average bandwidth usage of an access point in the MCC hotspot can be obtained from

$$\overline{B} = \sum_{c=1}^{C} \sum_{s=1}^{S} (1 - P_c) b_s. \quad (17)$$

- *Average reward of cloudlet:* The average reward of a MCC hotspot can be obtained from

$$\overline{R} = \sum_{\mathbf{x} \in \mathcal{X}} \sum_{a \in \mathcal{A}_{\mathbf{x}}} r(\mathbf{x}, a) \tau_{\mathbf{x}}(a) z^*_{\mathbf{x},a}. \quad (18)$$

IV. Numerical Results

We consider an MCC hotspot providing services for customers as depicted in Fig. 1. There are two classes of users (i.e., $C = 2$), namely, "member" and "nonmember". The "member" class (class-1) will have higher priority than that of "nonmember" class (class-2). Specifically, the blocking probability for member user class-1 has to be maintained less than or equal to 0.05. Furthermore, we assume that there are two types of service (i.e., $S = 2$) that the MCC hotspot provides, namely, service-1 and service-2. We also assume that the resource on the cloudlet is divided into 6 portions. In particular, a cloudlet can support maximum 6 virtual machines and each of which corresponds to a portion of resource (i.e., CPU time, memory, and storage space). If the service-1 is executed on the MCC hotspot, it will occupy 3 portions and 10 Mbps of bandwidth. If the service-2 is run on the MCC hotspot, it will consume 2 portions and 20 Mbps of bandwidth. The rest of parameters are set as follows. The departure rate for mobile users is equal to 1 request per minute. The arrival rate for mobile users is varied from 0.1 to 0.9 requests per minute.

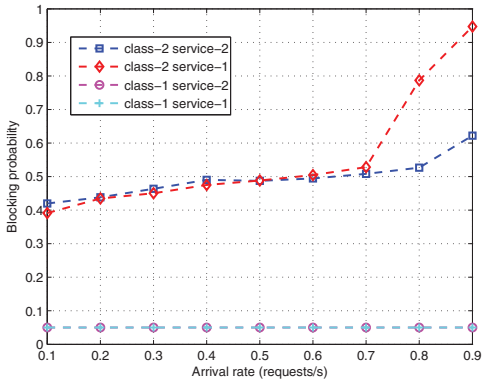

Fig. 2. Blocking probability of mobile users accessing MCC hotspot.

As depicted in Fig. 2, when we limit the blocking probability for a member user to be lower than or equal to 0.05, the probability that the request from a member user will be rejected is always smaller than 0.05. However, the blocking probability for a nonmember user will be high (above 0.4). If the arrival rate increases from 0.1 to 0.9, we can observe that the blocking probability for a nonmember user will increase steadily from 0.4 to 0.7. However, when an arrival rate is over 0.7, the blocking probability of a nonmember user goes up dramatically to 0.62 and 0.94 respectively. The reason here is because when the arrival rate of mobile users reaches 0.7, the bandwidth usage of the MCC hotspot reaches 30 Mbps, which is the maximum available bandwidth provided by an access point in MCC hotspot. In Fig. 3, if we do not have

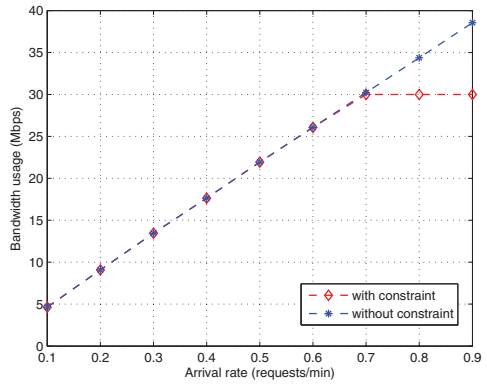

Fig. 3. Bandwidth usage of an access point in MCC hotspot.

the bandwidth constraint as defined in (13), when the arrival rate reaches 0.7, the bandwidth usage will be higher than 30 Mbps. We also observe that, with the constraint defined in (13), the server resource utilization of the cloudlet will be affected. As shown in Fig. 4, with the optimal policy obtained from solving SMDP, the utilization of the cloudlet is relatively high (always over 90%). When the arrival rate

increases, the utilization of cloudlet will be higher. However, when the arrival rate reaches 0.7 where at this point the total bandwidth usage reaches the maximum bandwidth, with the bandwidth constraint defined in (13), the utilization of the cloudlet remains constant approximately at 95.6%.

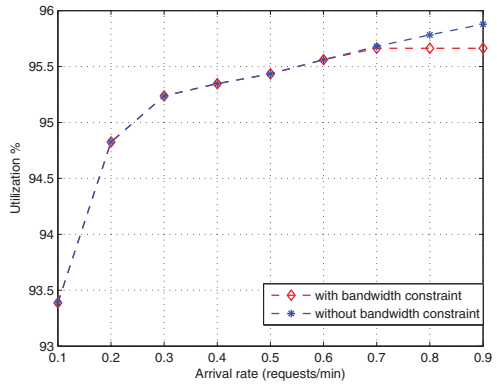

Fig. 4. Cloudlet utilization

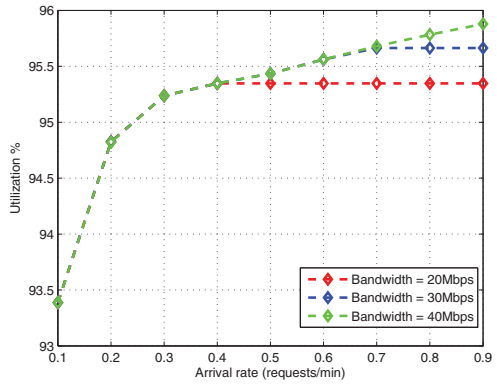

Fig. 5. The utilization of cloudlet when the bandwidth is varied.

Then, we consider one cloudlet and the bandwidth of the access point is varied. We can observe that the utilization of the cloudlet is always high. The utilization increases as the arrival rate increases as shown in Fig. 5. However, when the arrival rate increases too much (over 0.7), the utilization does not increase and maintains at a certain level since the system reaches the limit. The similar result is observed in Fig. 6 when the bandwidth is fixed at 30 Mbps and the size of cloudlet (i.e., the number of servers) is varied. In this case, the utilization of the cloudlet decreases when the number of servers increases. This means that the resource of cloudlet (e.g., three servers) is more than the demand from mobile users, especially when the bandwidth is fixed at 30 Mbps. In other words, the resource of cloudlet is wasted.

V. Summary

In this paper, we have developed a resource allocation model for the mobile cloud computing hotspot with a cloudlet. The optimization model based on semi-Markov decision process has been proposed. Our aim is to find an optimal policy in allocating resource of the bandwidth and cloudlet to meet the QoS requirements of mobile users to run mobile applications. With the optimal policy, it has been shown that the utilization of resource of cloudlet can be improved, while at the same time the users in different classes can be differentiated. In the future, we will extend this model by considering the cooperation among multiple MCC hotspots and also the resource sharing between cloudlet and the cloud.

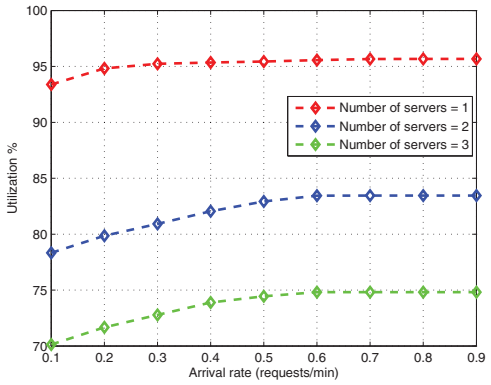

Fig. 6. The utilization of cloudlet when the size of cloudlet (i.e., the number of servers) is varied.

References

[1] D. B. Hoang, and L. Chen, "Mobile Cloud for Assistive Healthcare (MoCAsH)," in *Proceedings of the 2010 IEEE Asia-Pacific Services Computing Conference (APSCC)*, pp. 325, February 2011.

[2] R. Ferzli and I. Khalife, "Mobile cloud computing educational tool for image/video processing algorithms," in *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE)*, pp. 529, March 2011.

[3] Z. Ye, X. Chen, and Z. Li, "Video based mobile location search with large set of SIFT points in cloud," in *Proceedings of the 2010 ACM multimedia workshop on Mobile cloud media computing (MCMC)*, pp. 25-30, 2010.

[4] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches," *Wireless Communications and Mobile Computing (WCMC)*, accepted.

[5] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," IEEE Pervasive Computing, vol. 8, no. 4, pp. 14-23, October 2009.

[6] H. Xingye, L. Xinming, and L. Yinpeng, "Research on resource management for cloud computing based information system," in *Proceedings of International Conference on Computational and Information Sciences (ICCIS)*, pp. 491-494, December 2010.

[7] T. Lin and S. Wang, "Cloudlet-screen computing: A multi-core-based, cloud-computing-oriented, traditional-computing-compatible parallel computing Paradigm for the masses," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1805-1808, June 2009-July 2009.

[8] H. Liang, L. X. Cai, H. Shan, X. Shen, and D. Peng, "Adaptive resource allocation for media services based on semi-Markov decision process," in *Proceedings of International Conference Information and Communication Technology Convergence (ICTC)*, pp. 220 - 225, December 2010.

[9] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Hoboken, NJ: Wiley, 1994.

[10] A. Rudenko, P. Reiher, G. J. Popek, and G. H. Kuenning, "Saving portable computer battery power through remote process execution," *Journal of ACM SIGMOBILE on Mobile Computing and Communications Review*, vol. 2, no. 1, January 1998.