Intelligent Workload Management of Computing Resource Allocation for Mobile Cloud Computing

[1]Muzammil H Mohammed, [2]FaizBaothman
[1]*Assistant Professor, Department of Information Technology,*
College of Computers and Information Technology TaifUniversity, Taif, Saudi Arabia
Associate Professor, Department of Computer Science,
College of Computers and Information Technology, TaifUniversity, Taif, Saudi Arabia

Abstract: *Mobile cloud computing (MCC) allows mobile devices to source their computing, storage and alternative tasks onto the cloud to realize a lot of capacities and better performance. one in all the foremost important analysis problems is however the cloud will expeditiously handle the attainable overwhelming requests from mobile users once the cloud resource is proscribed. during this paper, a unique MCC adaptative resource allocation model is projected to realize the optimum resource allocation in terms of the greatest overall system reward by considering each cloud and mobile devices. to realize this goal, we have a tendency to model the adaptative resource allocation as a semi-Markov decision process (SMDP) to capture the dynamic arrivals and departures of resource requests. Intensive simulations square measure conducted to demonstrate that our projected model can do higher system reward and lower service obstruction likelihood compared to ancient approaches supported greedy resource allocation algorithmic program. Performance comparisons with numerous MCC resource allocation schemes are provided.*

Keywords**:** *Cloud computing, mobile cloud computing, semi-Markov decision process, QoS, cloud service supplier.*

I.INTRODUCTION

Cloud computing is a new computing service model with characteristics like resource on demand, pay as you go, and utility computing [1]. It provides new computing models for each service suppliers and individual customers, which may be loosely classified into infrastructure as a service (IaaS), platform as a service (PaaS), and Software as a service (SaaS). Good phones square measure expected to overtake PCs and become the foremost common net access entities worldwide as foreseen by Gartner [2]. Since mobile devices (MDs) have a lot of benefits like quality, flexibility, and sensing capabilities over mounted terminals, integration mobile computing and cloud computing techniques may be a natural and sure approach to make new mobile applications, that has attracted plenty of attention in each world and trade community. As a result, a replacement analysis field, known as mobile cloud computing (MCC), is rising.In [3], Huang et al. given a replacement MCC infrastructure, known as MobiCloud, wherever dedicated virtual machines (VMs) square measure appointed to mobile users to boost the protection and privacy capability. In such associate degree MCC setting, the system procedure resources, like mainframe, storage, and memory, square measure divided into many service provisioning domains supported the cluster geographical distribution. every domain consists of multiple VMs, and every VM handles elements of cloud computing resource (i.e., CPU, storage and memory, etc.). once the MCC service provisioning domain receives a service request from a mobile device, it must create a choice on whether or not to just accept the request; and the way abundant Cloud resources ought to be allotted if the request is accepted. though the Cloud resource is thought of as unlimited compared with the computing resource during a single mobile device, in apply, a geographically distributed cloud system typically contains restricted resource at an area service provisioning domain. once all the Cloud resources square measure occupied inside the native service provisioning domain, the service request from mobile device are going to be rejected or migrated to a nonlocal service provisioning domain owing to the resource inaccessibility. The rejection of a service request not solely degrades the user satisfaction level however conjointly reduces the system reward that is typically outlined as a metric that has the system profits and price.

The Cloud financial gain will increase with the amount of the accepted services. However, it's undoubtedly not true that cloud service supplier (CSP) would really like to acccept service requests as several as attainable, since a lot of accepted services occupy a lot of cloud resources, and a lot of probably a replacement request are going to be rejected once the network resource is proscribed, that degrades the QoS level of users. The rewards of the foremost existing Cloud resource allocation strategies solely contemplate the financial gain on behalf of the CSP. to get a comprehensive system reward of MCC, the client QoS and user satisfaction level ought to be taken under consideration within the system reward in addition. Therefore, our analysis goal is to deal with the subsequent questions: the way to get the greatest overall system rewards by taking under consideration from each the service supplier aspect and also the client aspect whereas satisfying a definite QoS level.

MCC resource allocation model supported semi-Markov call method (SMDP) to realize the target mentioned higher than. Our projected MCC model considers not solely the incomes of acceptive services, however conjointly the price resulted from VM occupation within the Cloud. Moreover, alternative factors together with service precessing time of each Cloud and MD battery consumption of mobile device are taken under consideration. Thus, the general economic gain is set by a comprehensive approach that considers all the factors mentioned higher than.The contributions and essence of this projected model square measure listed as follows.(i)semi-Markov decision process (SMDP) is applied to derive the optimum resource allocation policy for MCC.(ii)The projected model permits adaptative resource allocations, that is, multiple Cloud resources (i.e., the amount of VMs) is allotted to a service request supported the accessible Cloud resource within the service domain so as to maximise the resource utilization and enhance the user expertise. (iii)The greatest system rewards of Cloud is achieved by victimization the projected model and by taking into the concerns the expenses and incomes of each Cloud and mobile devices.

Recent analysis work for Cloud computing has shifted its focus from the Cloud for mounted user to Cloud for mobile devices [4], that allows a replacement model of running applications between resource-constrained devices and Internet-based Cloud. Moreover, resource-constrained mobile devices will source computation/ communication/ storage intensive tasks onto the Cloud. Clone Cloud [5] focuses on execution augmentation with less thought on user preference or device standing. Elastic applications for mobile devices via Cloud computing were studied in [6]. In [3], Huang et al. given associate degree MCC model that enables the mobile device connected operations residing either on mobile devices or dedicated VMs within the Cloud. [7] proposes some way victimization traffic-aware virtual machine (VM) placement to boost the network quantifiability by optimizing the location of VMs on host machines.Although resource management in wireless networks has been extensively studied there square measure few previous works that specialize in resource management of Cloud computing and particularly mobile cloud computing. In [11], associate degree economic mobile cloud computing model is given to determine the way to manage the computing tasks with a given configuration of the Cloud system. That is, the computing tasks is migrated between the mobile devices and also the Cloud servers. A game theoretical resource allocation model to apportion the Cloud resources per users' QoS necessities is projected in [12]. within the past few years, some analysis work centered on application of specific resource management in Cloud computing victimization virtual machines or finish servers in knowledge center. In [13], authors propose a replacement OS that allows resource-aware programming whereas allowing high-level reusable resource management policies for context-aware applications in Cloud computing. Lorincz et al. [14] address the matter of resource management in linguistics event process applications in Cloud computing. Tesauro et al. [15] propose a reinforcement learning based management system for dynamic allocation of servers trying to maximize the profit of the host data center in Cloud computing. In [16], Boloor et al. propose a generic request allocation and scheduling scheme to achieve desired percentile service level agreements (SLA) goals of consumers and to increase the profits to the cloud provider.

II. System Model

A major benefit of MCC over the traditional client-server mode is that MDs can have more capabilities and better performance (i.e., less processing time, energy saving, etc.) when they outsource their tasks onto the Cloud. The outsourcing procedure can be implemented by using weblets (application components) to link the services between the Cloud and the mobile devices. A weblet can be platform independent such as using Java or .Net bytecode or Python script or platform dependent, using a native code. Some research work [5] focuses on the algorithm to decide whether to offload the weblet from MD to the Cloud (i.e., run on one or more virtual nodes offered by an IaaS provider) or run the weblet on the MD itself. In this way, a mobile device can dynamically expand its capabilities, including computation power, storage capacity, and network bandwidth, by offloading an elastic application service to the Cloud. The choice made by mobile device on whether to offload the task onto the Cloud can refer to the mobile device's status such as CPU processing capability, battery power level, and network connection quality and security. The service scenario of the proposed model is the task offloading from MD onto the Cloud. Also, the task offloading procedure can be done in a way that MD sends a service request to the Cloud firstly, then the task is further offloaded to the Cloud once the service request is accepted by the Cloud.

As shown in Figure 1, a VM is responsible for managing the weblet's loading, unloading, and processing in the mobile Cloud. Each VM has the capacity to hold one weblet at a time for handling migrated weblet request, and two types of service requests are defined to be handled by a VM: (i) paid: a paid weblet service request is sent to the service provisioning domain from a mobile device; (ii) free: a free weblet service request is sent to the service provisioning domain from a mobile device. Figure 1 demonstrates the relationship between the paid/free service requests and the VMs of the service provisioning domain.

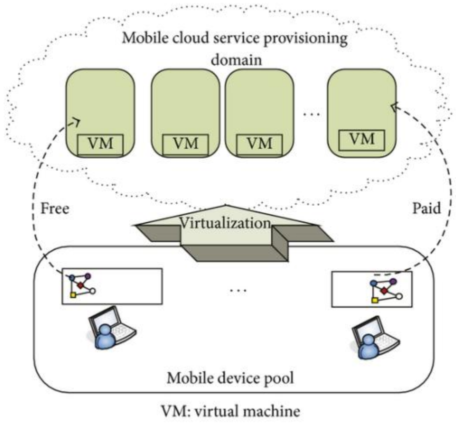

Fig.1: Reference model of mobile cloud computing.

The MCC service architecture is based on the MobiCloud framework presented in [3], in which a VM can handle a portion of Cloud system resources (CPU, memory and storage, etc.) that can satisfy the minimal resource requirement to process an application offloading service in the MCC system. Within the local MobiCloud service provisioning domain, the resource capacity, in terms of the number of VMs, is limited. Thus, if the demands of the arriving service requests exceed the number of available VM resources in a certain service domain, the following service requests will be rejected (or migrated to a remote service provisioning domain). On the other hand, if the demands of the arriving service requests are lower than the number of the available VMs, more VMs can be assigned to one service request to maximally utilize the Cloud resource and achieve a better performance and QoS. Our analytical model is based on a single local service domain. The analysis of local service migrations to remote service domains is regarded as the future study.

III. System Description

An MCC system chiefly consists of 2 entities, VM and physical MD. A VM is that the minimum set of resources which will be allotted to AN MD upon receiving its service request. Since AN MD could be a wireless node with restricted computing capability and energy provide, it will source its mobile codes (i.e., weblet) of AN application service to the Cloud. Then, the Cloud can decide variety of VMs to be allotted to the arrival service request if the choice for the service request created by the Cloud is accepted.

we take into account a service provisioning domain with VMs. the most range of VMs which will be allotted to a Cloud service is VMs ,where typically, the length for running a mobile application service within the Cloud depends on the amount of VMs allotted thereto service. the link between the interval of AN application service and also the range of allotted VMs within the Cloud may be expressed as a perform denoted as . Assume that the time to method AN application service by exploitation one VM during a service provisioning domain is , so the time to handle the service is that if VMs ar allotted thereto service. the upper computing speed for AN application service during a service provisioning domain means that the upper user satisfaction level, that is that the major a part of the entire system reward of the Cloud. Thus, so as to boost the entire system reward of a service provisioning domain by increasing the user satisfaction level, the normal greedy formula [17] invariably decides to assign highest VMs to the service. however on the opposite hand, if the Cloud computing resources (denoted by the amount of VM) allotted to this service by the service provisioning domain ar too high, then the subsequent many arrival service requests could also be rejected by the service provisioning domain thanks to light on the market Cloud computing resources, that decease the user satisfaction level. As a result, the system rewards of that MCC service provisioning domain degrade in addition.It may be additional difficult once we take into account each the rewards and prices of mobile devices. price concerned within the MD facet mustn't be neglected, which suggests that the entire system reward ought to take into account not solely the rewards of the mobile Cloud itself, however additionally the incomes and also the prices of MD, like the saved battery energy if the service is processed within the mobile Cloud and also the expense of the battery energy and also the interval of MD if the applying service is processed on the MD domestically. To model this complicated dynamic MCC resource allocation method, while not loss of generality, we have a tendency to assume that the arrival rates of each paid and free service requests follow Poisson distributions with mean rate of and , severally. The life time of services follows exponential distributions. The mean holding time of a service that is allotted only 1 VM within the service provisioning domain is . Thus, the holding time of the service allotted VMs within the domain is, which means that the mean departure rate of finished service.

Since the choice creating epoch is haphazardly generated within the system, we have a tendency to use semi-Markov call method (SMDP) to model the dynamic MCC resource allocation method supported the system description we have a tendency to bestowed on top of. SMDP could be a random dynamic programming technique, which might be wont to model and solve optimum dynamic deciding issues. There ar six following parts within the SMDP model: (a) system states; (b) action sets; (c) the events that cause the calls; (d) decision epoches; (e) transition probabilities; and (f) reward. Within the following, we have a tendency to 1st gift the system states, the actions, the events, and also the reward model for the MCC system.

IV. System States

According to the belief, there are total VMs in one service provisioning domain, and VM may be allotted to the service request, that is from one to,where. However, the arrival of paid application services request and free application service request and also the departure of the finished service are distinct events. Thus, the system states may be delineated by {the range theamount the quantity} of the running Cloud services that occupy constant number of VMs and also the events (including each arrival and departure events) within the service provisioning domain. Here, we have a tendency to use to point the amount of VMs allotted to at least one application service (denoted as allocation theme as bestowed in Section three.1), Therefore, the amount of the running Cloud services that occupy VMs in one service provisioning domain may be denoted within the MCC system model, we are able to outline 2 varieties of service events: a paid or free service request arrives from AN MD severally and also the departure of a finished application service occupying VMs within the current service provisioning domain, Thus, the event within the MCC system may be delineated and also the system state may be expressed .

V. Actions

For a system state of the service provisioning domain with AN incoming service request from AN MD the mobile Cloud has to build a call on whether or not to just accept the service request and what's the allocation theme, if the choice is acceptance. If the choice is acceptance, then the allocation theme is assigned to the arrival service request; therefore, the action to assign the allocation theme may be denoted as. whereas if the choice is rejection supported the entire system reward, which suggests no VM are going to be assigned , therefore the paid or free service request are going to be rejected and also the application can run on the MD itself. Then, the action to reject the service request may be denoted for the departure of a finished service within the service provisioning domain (i.e.,), the action for this event may be thought-about on calculate this on the market Cloud resources .Based on the system state and its corresponding action, we are able to appraise the entire mobile Cloud system reward that is computed supported the financial gain and also the price as follows: wherever is that the internet payment financial gain for the Cloud and MDs and denotes the system price.

The net payment financial gain ought to take into account the payment from MD to the mobile Cloud, the saved battery energy of MD, and also the consumed time of mobile Cloud to method the service if the service is run within the mobile Cloud, the consumed battery energy, and also the consumed time of MD if the service is run on MD domestically.Thus, cyber web payment financial gain is computed the service provisioning domain obtained from the MD once it accepts a paid service request from the MD. denotes the time consumed on sending the service request from MD to the service provisioning domain through wireless affiliation, whereas denotes the worth per unit time, that has constant measuring unit because the financial gain. Thus, denotes the expense measured by the time consumed on sending the service request from MD to the service provisioning domain. represents the expense measured by the battery energy consumed by the MD once the service request is rejected by the service provisioning domain and run on the MD domestically, that has constant measuring unit because the financial gain. is that the weight issue that satisfies . Let denote the time to method AN application service by exploitation one mobile device, then represents the expense measured by the time consumed to method the applying exploitation one mobile device. Similarly, denotes the expense measured by the time consumed to method the service exploitation one VM during a service provisioning domain. Therefore, denotes the expense measured by the time consumed to method the service exploitation VMs during a service provisioning domain.

VI. SMDP-Based Mobile Computing Model

Based on the SMDP model, we've already outlined the system states, action sets, the events, and reward for the MCC system within the last section, then we'd like to outline the choice epoches and procure the transition chances to calculate the most semipermanent whole system reward.There ar 3 varieties of events within the MCC system (i.e., AN arrival of a paid service request, AN arrival of a free service request, and a departure of a finished service). Consequent call epoch happens once any of the 3 varieties of events takes place. supported our assumption, the arrival of service request follows distribution and also the departure of finished service follows exponential distribution. Thus, the expected time length between 2 call epoches follows exponential distribution in addition. Then, the mean rate of expected time may be painted because the expected discounted reward throughout may be obtained supported the discounted reward model outlined in [18, 19], wherever could be a continuous-time discounting issue and outlined in (4), (6), and (7), severally.

Then the sole component left to be calculated is that the transition chances. To calculate the transition chances, we have a tendency to show AN example in Figure two.

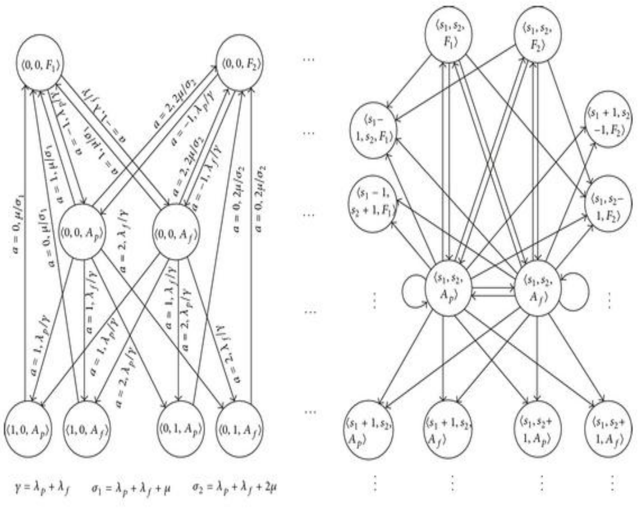
$\langle 0,0,F_1\rangle$
$\langle 0,0,F_2\rangle$
$\langle 0,0,A_p\rangle$
$\langle 0,0,A_f\rangle$
$\langle 1,0,A_p\rangle$
$\langle 1,0,A_f\rangle$
$\langle 0,1,A_p\rangle$
$\langle 0,1,A_f\rangle$
$a = 0, \mu/\sigma_1$
$a = 0, \mu/\sigma_1$
$a = 1, \mu/\sigma_1$
$a = -1, \lambda_p/\gamma$
$a = 1, \mu/\sigma_1$
$a = -1, \lambda_f/\gamma$
$a = 2, 2\mu/\sigma_2$
$a = -1, \lambda_p/\gamma$
$a = 2, 2\mu/\sigma_2$
$a = -1, \lambda_f/\gamma$
$a = 0, 2\mu/\sigma_2$
$a = 0, 2\mu/\sigma_2$
$a = 2, \lambda_f/\gamma$
$a = 1, \lambda_p/\gamma$
$a = 1, \lambda_f/\gamma$
$a = 1, \lambda_p/\gamma$
$a = 2, \lambda_p/\gamma$
$a = 1, \lambda_f/\gamma$
$a = 2, \lambda_p/\gamma$
$a = 2, \lambda_f/\gamma$
$\gamma = \lambda_p + \lambda_f$
$\sigma_1 = \lambda_p + \lambda_f + \mu$
$\sigma_2 = \lambda_p + \lambda_f + 2\mu$
$\langle s_1, s_2, F_1\rangle$
$\langle s_1, s_2, F_2\rangle$
$\langle s_1 - 1, s_2, F_1\rangle$
$\langle s_1 - 1, s_2 + 1, F_1\rangle$
$\langle s_1 + 1, s_2 - 1, F_2\rangle$
$\langle s_1, s_2 - 1, F_2\rangle$
$\langle s_1, s_2, A_p\rangle$
$\langle s_1, s_2, A_f\rangle$
$\langle s_1 + 1, s_2, A_p\rangle$
$\langle s_1 + 1, s_2, A_f\rangle$
$\langle s_1, s_2 + 1, A_p\rangle$
$\langle s_1, s_2 + 1, A_f\rangle$

Fig. 2

In this example, without loss of generality, we assume that there are only two allocation schemes, which means,the transition probabilities of allocation schemes can be deduced. Let denote the state transition probability from the current state to the next state when action is chosen. Then, the transition probability can be expressed as following.

For the states , the action for this departure state is always which means , then the transition probability can be obtained as where , the maximal long-term discounted reward is obtained based on the discounted reward model defined in [18, 19] and can be denoted as where , and can be obtained in (8), (9), (10), and (11).In the reward equation (8), the first part is that the revenue is a lump earnings of the reward and the second part is that the cost is a continuous-time payment of the reward. Thus, the reward function needs to be uniformized to obtain the uniformized long-term reward, then the discrete-time discounted Markov decision process can be used in this model. Based on the assumption 11.5.1 in [19], we need to find a constant satisfying to obtain the uniformized long-term reward by utilizing (11.5.8) in [19]. denote the uniformized transition probability, the long-term reward, and the reward function, respectively.

VII. Performance Analysis

The likelihood of allocation theme , that is outlined because the likelihood that VMs area unit allotted for a cloud service, is a very important performance metric for guaranteeing the user satisfaction level and therefore the Cloud resource utilization quantitative relation. it's terribly helpful for the operator to manage the system capacity/utilization standing supported the system parameters of the service provisioning domain like arrival rate, departure rate, and therefore the VM variety of Cloud resource. Meanwhile, interference service request doesn't solely mean the loss of whole system reward, however conjointly means that the degradation of users' satisfaction level. Then, the interference likelihood, that is that the likelihood that interference the cloud service requests from mobile device, is another necessary performance metrics for the service provisioning domain. From the reward operate (18) and likelihood equations (14), (15), and (16), the expected total discounted reward at state is said with the arrival rates of paid service request and free service request , the departure rate of every allocation theme, the occupied Cloud resource expressed by the quantity of being occupied VMs and therefore the capability of the service provisioning domain (i.e., the overall variety of VMs-). For a given service provisioning domain associate degreed a precise system state of an arrival of service request theon top of parameters area unit fastened. As a result, the steady-state likelihood of every state may be obtained from the likelihood equations (14), (15), and (16). Thus, the possibilities of every allocation theme and interference likelihood also can be achieved through the steady-state likelihood of every state.Let denote the steady-state likelihood of the system state within the service provisioning domain. From the instance in Figure a pair of , the steady-state likelihood of may be classified as 3 types: the arrival of a paid service request; the arrival of a free service request; the departure of a finished service with allocation theme. supported the likelihood equations (14), (15), and (16), the steady-state chances and may be derived as follows ,the parameters determined by the correlative actions severally . Similarly, the steady-state likelihood may be earned as wherever and area unit outlined by the connected actions severally as total of the steady-state chances for all states equals to one.

Therefore, the steady-state likelihood of every state in associate degree MCC service provisioning domain may be obtained by resolution (19), (20), (22), and (24). Thus, as a result, for the service request arrival states in one service provisioning domain, the likelihood of every action may be achieved, that is that the quantitative relation of the total of all steady-state chances with constant action to the total of the steady-state chances of all service request arrival states in one domain. Let and denote the likelihood of every action for paid service request and free service request, severally, then, and may be expressed as supported (26) and (25), the interference likelihood for the service request arrival states (i.e, and) in one service provisioning domain may be obtained and denoted as and , severally.

The high values not solely mean the loss of the total system reward however conjointly the decrease of the QoS of the service provisioning domain. Thus, the interference chances and area unit vital metrics to live the potential and QoS of a service provisioning domain.within the next section, we'll illustrate the relationships between the interference likelihood and therefore the parameters supported the simulation results.

The performance of the planned economic MCC model supported SMDP by victimization an incident driven machine compiled by Matlab [20] and compare our planned model with the normal greedy algorithmic program. Since the paid service demands a better QoS level compared with alternative free services, so our simulation in the main focuses on the performance of paid service.In our simulation, the peak variety of VMs is , and therefore the theme that allocates VMs to a service is denoted as allocation theme . The time to method associate degree application service by the Cloud is assumed as a linear operate of the quantity of VMs allotted to the service, which might be denoted as . Thus, the worth will be obtained because the total resource capability of the service provisioning domain is up to VMs. Unless otherwise such, the arrival rates of the paid and free service request area unit severally and therefore the departure rate of finished service occupying one VM.. Since the time to method the appliance service occupying one VM is that the departure rate of finished service occupying multiple VMs is that is delineated in Section three. Thus, the departure rates of finished service occupying one, two, and 3 VMs area unit severally. To assure reward computation convergence the continuous-time discounting issue is about to the simulation results area unit collected with every experiment running s, and every experiment runs rounds.

VIII. Optimum Actions

The actions of optimum resource allocation at every system state with completely different arrival rates of the paid service. The numbers within the tables represent the optimum choices created on state. Once no user is within the service provisioning domain, three VMs (which implies that the action is made) area unit allotted to the paid service in each 2 eventualities, once a paid service request arrives. If there area unit services within the service provisioning domain, which suggests that the quantity of the occupied VMs is , thus, there area unit unoccupied VMs obtainable within the service provisioning domain. Our planned model allocates VMs to the paid service request once the arrival rate of paid service requests is low and allocates VMs to the paid service request once the arrival rate of paid service requests is high , which suggests that once the arrival rate of paid service requests will increase, our model becomes additional conservative to allot resources to the paid service requests. the rationale is, for the state , the corresponding lump incomes as a result of the little variance between the lump incomes obtained by allocating and VMs to the paid service request, once the arrival rate of paid service requests will increase ,our model prefers action apart from action , since action will accommodate additional paid services to achieve higher rewards of the MCC system than action , that consumes additional Cloud resources of the service provisioning domain.

To evaluate the performance of the planned dynamic resource allocation model, we tend to compare the long-run reward and interference likelihood of the paid service between our model and greedy technique in Figures three, 4, and 5. In Figure three, the reward of paid service of our model will increase at the start, then falls down with the rise of the arrival rate of paid service requests , whereas the reward of paid service victimization the greedy technique declines continuously. It may be seen during this figure that the reward of the paid service of our planned model performs far better than that of greedy technique. In Figure four, with the rise of the arrival rate of the paid service requests, our model would rather to allot additional and VMs to the paid service request alternative VM; so, the dropping likelihood of our model is less than that of the greedy technique which might be seen in Figure five additionally. because the rejection has additional impact on the system lump financial gain compared with acceptance (in our simulation, the lump financial gain or fine of rejection is , whereas the corresponding lump incomes and so the lower the dropping likelihood of our model gains additional rewards of paid service than the greedy technique. we will conjointly see in Figure four that once the arrival rate of the paid service requests is over , the possibilities to allot and VMs (especially the likelihood of VM) exceed the likelihood to allot VM, that explains the rationale why the reward of paid service of our planned model falls down once the arrival rate of paid service requests exceeds as shown in Figure three. In a word, our model can do higher reward of paid service whereas keeping lower dropping likelihood of paid service requests at constant time examination with the greedy technique, that area unit shown in Figures three and five, severally. Thus, our model outperforms the greedy technique with the rise of arrival rate of paid service requests.

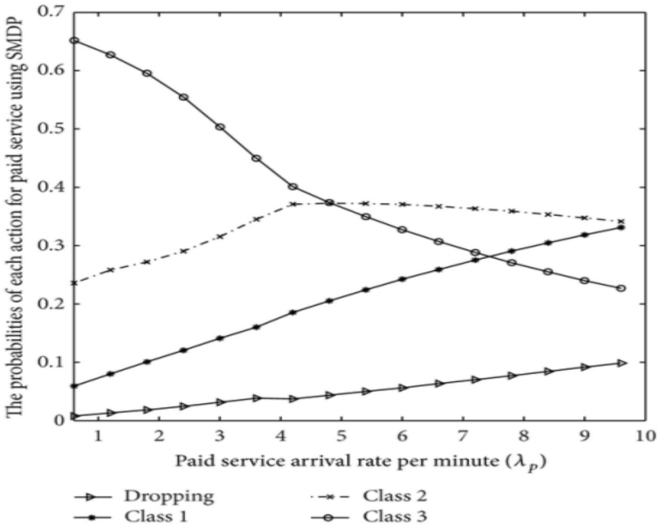

Fig.3

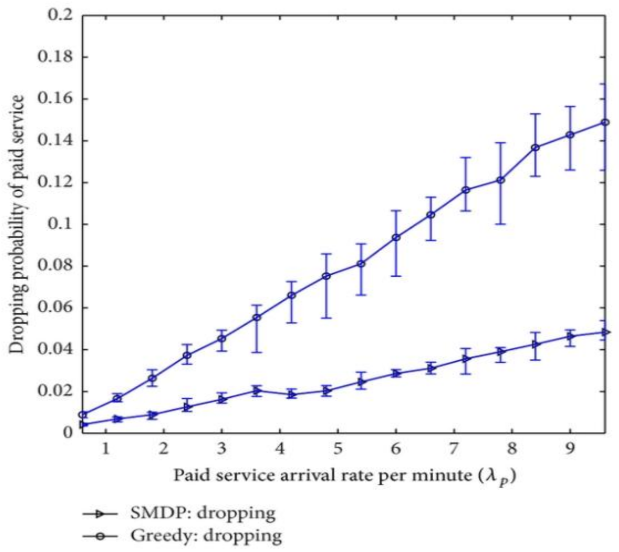

Fig.4

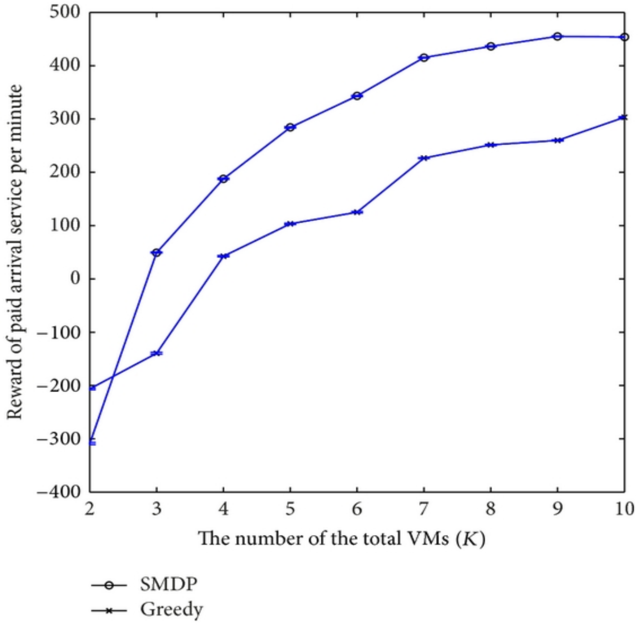

Fig 5

When the number of VMs () is less than , the rewards of both our model and greedy method are negative. This is because the absolute value of rejection cost () is much higher than the net lump rewards of acceptance in our simulation.When the number of total VMs in the service provisioning domain is low (and), the rejection probability of paid service requests is as high which results in the negative rewards for both our model and greedy algorithm. We also observed that when is less than , the reward of paid service of our model is lower than that of the greedy method.

The reason is that our model does not only consider the instant and future long-term income but also the cost of resource occupation of all running services in the service provisioning domain when deciding to allocate the Cloud resources to the paid service request, while the greedy method only considers the current income of paid service of the service provisioning domain. Then, when the Cloud resource of the service provisioning domain is less than VMs, our model is more conservative than the greedy method to allocate Cloud resources to the paid service request.

In Figure 5, we can also see that when the number of VMs () is less than , the reward of paid service of our model increases rapidly with the increase of , while when is greater than , the reward of paid service of our model increases slowly with the increase of , which implies that when the Cloud resource of the service provisioning domain exceeds the threshold, for the given arrival rate and departure rate, it has limited impact to increase the reward of paid service through increasing the Cloud resource of the service provisioning domain. Comparing the rewards of paid service between our model and the greedy method in Figure 5, it can be seen that our model outperforms over averagely than the greedy method. Meanwhile, as shown in Figure 5, the dropping probability of paid service requests of our model is lower than that of the greedy method over averagely as well, which proves that our model performs better than the greedy method with the increase of the total number of VMs (or Cloud resources) of the service provisioning domain as well.

Figure 6 shows the total rewards (rewards of paid service plus free service) of different arrival rates of free service requests of our proposed model, varying with the increase of arrival rate of paid service requests in the service provisioning domain. It can be seen that when the values of the arrival rates between paid service request and free service request are comparable, the total reward of our model increases with the increase of arrival rate of free service requests. On the other hand, when the arrival rate of free service requests is much larger than that of paid service requests, the total reward decreases rapidly, which results from the large increase of the arrival rate of free service requests which may cause more rejections for the following service requests.

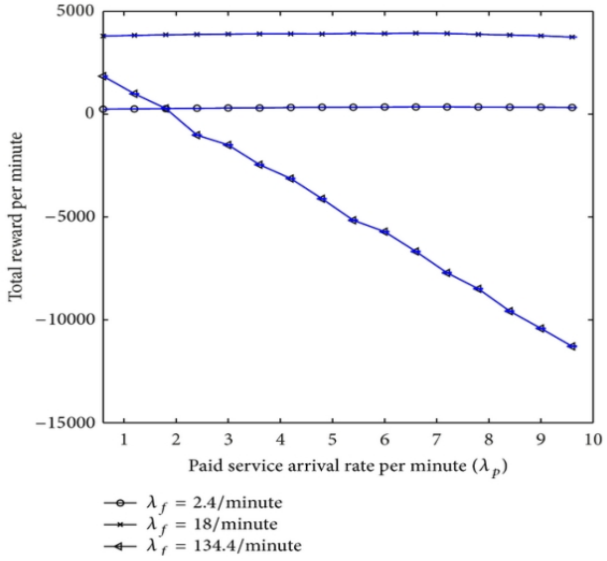

Fig.6

XI. Conclusion

In this paper, we propose an SMDP-based model to adaptively allocate Cloud resources in terms of VMs based on requests from mobile users. By considering the benefits and expenses of both Cloud and mobile devices, the proposed model is able to dynamically allocate different numbers of VMs to mobile applications based on the Cloud resource status and system performance, thus to obtain the maximal system rewards and to achieve various QoS levels for mobile users. We further derive the Cloud service blocking probability and the probabilities of different Cloud resource allocation schemes in our proposed model. Simulation results show that the proposed model can achieve a higher system reward and a lower service blocking probability compared with the traditional greedy resource allocation algorithm. In the future, we will study a more complex decision making model with different types of mobile application services, for example, the mobile application services which require different serving priorities. We will also investigate the optimal Cloud resource planning by determining the minimal Cloud network resources to achieve the maximal system rewards under given QoS constraints.

X. References

[1] M. Armbrust, A. Fox, R. Griffith, et al., “Above the clouds: a berkeley view of cloud computing,” Tech. Rep. UCB/EECS-2009-28,EECS Department, University of California, Berkeley, Calif, USA, 2009.

[2] M. Walshy, “Gartner: Mobile to outpace desktop web by 2013,” Online Media Daily.

[3] D. Huang, X. Zhang, M. Kang, and J. Luo, “Mobicloud: a secure mobile cloud frame-work for pervasive mobile computing and communication,” in Proceedings of 5th IEEE

International Symposium on Service-Oriented System Engineering, 2010.

[4] X. H. Li, H. Zhang, and Y. F. Zhang, "Deploying mobile computation in cloud service," in Proceedings of the 1st International Conference for Cloud Computing (CloudCom '09), p. 301, 2009.

[5] B. Chun and P. Maniatis, "Augmented smartphone applications through clone cloud execution," in Proceedings of the 12th USENIX HotoS, 2009.

[6] X. Zhang, J. Schiffman, S. Gibbs, A. Kunjithapatham, and S. Jeong, "Securing elastic applications on mobile devices for cloud computing," in Proceedings of the ACM workshop on Cloud Computing Security, pp. 127–134, 2009.

[7] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," in Proceedings of the IEEE INFOCOM, San Diego, Calif, USA, March 2010.

[8] L. X. Cai, L. Cai, X. Shen, and J. W. Mark, "Resource management and QoS provisioning for IPTV over mmWave-based WPANs with directional antenna," ACM Mobile Networks and Applications, vol. 14, no. 2, pp. 210–219, 2009.

[9] H. T. Cheng and W. Zhuang, "Novel packet-level resource allocation with effective QoS provisioning for wireless mesh networks," IEEE TransacTions on Wireless Communications, vol. 8, no. 2, pp. 694–700, 2009.

[10] L. X. Cai, X. Shen, and J. W. Mark, "Efficient MAC protocol for ultra-wideband networks," IEEE Communications Magazine, vol. 47, no. 6, pp. 179–185, 2009.

[11] H. Liang, D. Huang, and D. Peng, "On economic mobile cloud computing model," in Proceedings of the International Workshop on Mobile Computing and Clouds (MobiCloud '10), 2010.

[12] G. Wei, A. V. Vasilakos, Y. Zheng, and N. Xiong, "A game-theoretic method of fair resource allocation for cloud computing services," The Journal of Supercomputing, vol. 54, no. 2, pp. 252–269, 2009.

[13] K. Lorincz, B. R. Chen, J. Waterman, G. Werner-Allen, and M. Welsh, "Resource aware programming in the pixie os," in Proceedings of the SenSys, Raleigh, NC, USA, November 2008.

[14] K. Lorincz, B. Chen, J. Waterman, G. Werner-Allen, and M. Welsh, "A stratified approach for supporting high throughput event processing applications," in Proceedings of the DEBS, Nashville, Tenn, USA, July 2009.

[15] K. Boloor, R. Chirkova, Y. Viniotis, and T. Salo, "Dynamic request allocation and scheduling for context aware applications subject to a percentile response time sla in a distributed cloud," in Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science, Indianapolis, Ind, USA, November 2010.

[16] R. Ramjee, D. Towsley, and R. Nagarajan, "On optimal call admission control in cellular networks," Wireless Networks, vol. 3, no. 1, pp. 29–41, 1997.

[17] S. O. H. Mine and M. L. Puterman, Markovian Decision Process, Elsevier, Amsterdam, The Netherlands, 1970.

[18] M. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, New York, NY, USA, 2005.MathWorks, "Matlab," http://www.mathworks.com/.

[19] HongbinLiang,State Key Laboratory of Information Security, Institute of Information Engineering, The Chinese Academy of Sciences, Beijing 100093, China

[20] Tianyi Xing, School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, Sichuan 610031, China

[21] Lin X. Cai Arizona State University, 699 S Mill Avenue, Suite 464, Tempe, AZ 85281, USA