Optimal Allocation of Virtual Machines for Cloud-based Multimedia Applications

Xiaoming Nan, Yifeng He, Ling Guan

Department of Electrical and Computer Engineering
Ryerson University, Toronto, Ontario, Canada

Abstract—With the emergence of cloud computing, cloud-based multimedia applications have been increasingly adopted in recent years. There are two major challenges for multimedia application providers: the round trip time (RTT) requirement and the resource cost. In this paper, we study the virtual machine (VM) allocation problem for multimedia application providers to minimize the resource cost under RTT requirements. Specifically, we propose the optimal VM allocation schemes for single-site cloud and multi-site cloud, respectively. Moreover, we propose the greedy algorithms to efficiently allocate VMs in each case. Simulation results demonstrate that the proposed optimal VM allocation schemes can optimally allocate VMs to achieve a minimal resource cost.

I. INTRODUCTION

As the emerging computing paradigm, cloud computing manages a shared pool of servers in data centers to provide on-demand computing resources (e.g., computing power, platform, applications, etc.) as accessible services for users via the Internet. To enable the resource provisioning efficiently, the virtualization technology is applied to package the required CPU, memory, and storage resources into virtual machines (VMs). By using VMs, cloud resources can be provisioned or released with the minimal efforts. According to the service provisioning at different levels, three cloud service models have been proposed [1], namely Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), among which the SaaS model is the most familiar to individual cloud users.

In the SaaS cloud model, three major entities are involved in the service provisioning: cloud providers, application providers, and users. As the computing resource suppliers, cloud providers operate the infrastructure of data centers and provide VMs as service. By renting VMs from cloud providers, application providers can deliver different applications to users. As the consumers in the SaaS model, users send requests for interested applications and receive results from application providers. All three entities can benefit from the SaaS model. Users don’t need to install applications or buy software licences, application providers don’t have to make investment on servers, and cloud providers can acquire profits from VMs rental.

Among various cloud-based applications, multimedia applications strongly need assistance from cloud computing due to the compute-intensive and delay-sensitive requirements. For multimedia application providers, there are two major challenges: the round trip time (RTT) and the resource cost. The RTT in cloud is defined as the sum of the forward transmission delay from the user to the data center, the backward transmission delay from the data center to the user, and the service response time in the data center. Considering the delay-sensitive requirement of multimedia, the RTT is taken as a significant Quality of Service (QoS) factor to measure the performance of cloud-based multimedia applications. But it is a challenge to satisfy the RTT requirements of different multimedia applications for users in different regions. Besides the RTT, the resource cost is also a challenge. Generally, cloud providers can offer two VM pricing schemes: the reservation scheme and the on-demand scheme. Price of VMs in the reservation scheme is cheaper than that in the on-demand scheme. But in the reservation scheme, application providers have to subscribe a certain amount of VMs in advance for future usage. During the service provisioning, if the reserved VMs cannot meet the resource demands, the on-demand VMs can be rent instantly to meet the extra demands at the expense of a higher price rate. Facing different applications and resource demands, it is challenging for application providers to optimally allocate VMs for each application to achieve the minimal resource cost and satisfy the RTT requirements.

In this paper, we investigate the VM allocation problem for multimedia application providers to minimize the resource cost under the RTT requirements. Our contributions in this paper can be presented as follows. We propose the optimal VM allocation schemes for single-site cloud and multi-site cloud, respectively. In each case, we formulate and solve the optimal VM allocation problem to minimize the resource cost under the RTT constraints. Since the formulated optimization problem is an NP-hard problem, we propose the greedy algorithm in each case to efficiently allocate VMs. The proposed greedy algorithm is sub-optimal solution, which is demonstrated to perform close to the optimal solution by the simulation results.

II. RELATED WORK

Cloud computing, as the emerging computing paradigm, has widely attracted attentions of researchers from both industry [2], [3], [4] and academia [5-14]. Survey works on the general definition, challenges, and service models of cloud computing can be found from [5], [6]. Among different cloud service models, SaaS model has been identified as the most profitable model. According to the Gartner Group estimate [7], the worldwide SaaS sales reached $12.1 billion in 2011. The major cloud providers, including Amazon EC2 [2], Microsoft Azure
As multimedia processing has the delay-sensitive characteristic, the QoS requirement becomes crucial to the cloud based multimedia applications. Zhu et al. [8] present the concept of multimedia cloud and propose the media edge cloud computing architecture aiming to reduce the transmission delays. Ye et al. [9] employ the on-demand cloud resource to provide a video based location search service for mobile users. Wu et al. [10] integrate the cloud server with P2P streaming together to support the video on demand application. But in [8]-[10], they do not provide the theoretical analysis and optimal VM allocation for multimedia application providers.

The resource allocation of cloud has been studied in [11]-[14]. Lin et al. [11] propose a self-organizing model to manage cloud resources in the absence of centralized management control. Nan et al. [12] present the queuing model based optimal resource allocation for multimedia cloud. Chaisiri et al. [13] propose an optimal VM placement algorithm based on the stochastic integer programming. Nan et al. [14] present the optimal cloud resource allocation in priority service scheme to minimize the resource cost. However, they [11]-[14] do not consider the delay requirement of the multimedia applications. Compared to the work in [11]-[14], our paper is different in the following senses: 1) we study the resource allocation problem from the multimedia application providers’ perspective; 2) we take the RTT as the QoS factor for the cloud based multimedia applications; 3) we propose the theoretical optimal solution and practical greedy algorithm for multimedia application providers.

III. SYSTEM MODELS

In this section, we present our system models including the SaaS architecture, the pricing scheme, and the proposed VM allocation model.

A. SaaS Architecture

In cloud computing environment, the data center consists of a number of computing servers, which support the running and provisioning of VMs. Application providers rent a collection of different classes of VMs for usage. The different classes of VM instances generally have different configuration levels in terms of computing units, CPU frequency, memory size, and I/O rate. A number of VMs in the same class can work together as the virtual cluster [15] to provide the more powerful computing resources.

The SaaS architecture for multimedia application providers is illustrated in Fig. 1. The service process can be presented as follows. When the users’ requests arrive at data center, the dispatcher will distribute the requests to the corresponding application. In each application, different classes of VMs are allocated to provide service. The workload monitor performs a live monitoring on the type and number of requests, and forwards them to the Service Level Agreement (SLA) negotiator and the optimal VM allocation scheme. Based on the request arrival rate and the price rates, the SLA negotiator will determine the upper bound of the RTT for each application. Given the request arrival rate and the RTT constraints, the proposed optimal VM allocation scheme will optimally allocate the required number of VMs for each application to achieve the minimal resource cost. If the initially reserved VMs cannot satisfy the resource demands, the resource broker will request additional on-demand VMs from cloud providers by sending the VMs classes and amount specification. With the allocated on-demand VMs, the RTT requirement of each application can be guaranteed.

B. Pricing Scheme

There are two pricing schemes for VMs rental, i.e. the reservation scheme and the on-demand scheme. In the reservation scheme, the application providers have to make a one-time upfront payment for a VM instance and in turn receive a significant discount on the hourly rate for that instance. In the on-demand scheme, the VMs can be paid by the hourly usage without long-term commitment or upfront payment. But the price rate in the on-demand scheme is much higher than that in the reservation scheme. In each plan, the price rate is defined in dollars ($) per VM instance, and different classes of VMs have different price rates.

C. VM Allocation Model

We propose the VM allocation model to study the VM allocation problem. Since the workload is time varying, the cloud resources have to be allocated dynamically. Therefore, we divide the time domain into time slots. In our model, we will denote by \(t\) the time slot for VM allocation. Suppose that \(N\) classes of VM configurations are supplied by the cloud provider. Let \(p^{res}\) and \(p^{dem}\) denote the price rates of the class-i \((\forall i = 1, 2, \ldots, N)\) VM instance in reservation scheme and on-demand scheme, respectively. In the initial reservation phase, the application providers will reserve a certain number of different classes of VMs at the reservation price rates, and let \(R_i^{res}\) denote the number of initially reserved class-i \((\forall i = 1, 2, \ldots, N)\) VMs. Suppose that there are \(M\) types of applications provided. For each type of application, different classes of VMs have different service rates. Let \(\mu_{ij}\) be the service rate of the class-i \((\forall i = 1, 2, \ldots, N)\) VM instance for processing the type-j \((\forall j = 1, 2, \ldots, M)\) application requests.
Suppose that the users who request the cloud-based multimedia services can be divided into \( R \) groups, where the group of users can be defined as a set of users located in the same geographic area. This allows us to assume that the forward and backward transmission delays of users in the same group are identical. The similar assumption is made in [16], [17]. Let \( T_{fud}(t) \) and \( T_{bad}(t) \) denote the forward and backward transmission delays in time slot \( t \) for users in group \( r \) (\( \forall r = 1, 2, \ldots, R \)), respectively. We assume that the arrivals of requests from group-\( r \) users for type-\( j \) application in time slot \( t \) follow a Poisson Process with mean arrival rate of \( \lambda_{ij}^{(t)} \). According to the composition property of Poisson distribution [18], the total request arrivals for the type-\( j \) application also follow a Poisson Process with the mean arrival rate of \( \lambda_{ij}^{(t)} = \sum_{r=1}^{R} \lambda_{ij}^{(t)} \).

Let \( K_{ij}^{res(t)} \) and \( K_{ij}^{dem(t)} \) denote the number of class-\( i \) VMs allocated to type-\( j \) application in time slot \( t \) in the reservation scheme and on-demand scheme, respectively. We will propose the optimal VM allocation schemes to determine the optimal value of \( K_{ij}^{res(t)} \) and \( K_{ij}^{dem(t)} \) (\( \forall i = 1, 2, \ldots, N, \forall j = 1, 2, \ldots, M \)) to minimize the resource cost under the RTT constraints.

### IV. OPTIMAL VM ALLOCATION SCHEME FOR SINGLE-SITE CLOUD

In this section, we study the optimal VM allocation scheme for single-site cloud, in which the application providers deploy multimedia applications at the single data center. Thus, all requests are sent to the data center for service.

Our objective is to minimize the resource cost for application providers. As presented in Section III, the cost of allocated VMs in time slot \( t \) can be formulated as \( C_{\text{sum}}^{(t)} = \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij}^{res} K_{ij}^{res(t)} + \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij}^{dem} K_{ij}^{dem(t)} \), where \( \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij}^{res} K_{ij}^{res(t)} \) is the cost of VMs in the reservation scheme and \( \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij}^{dem} K_{ij}^{dem(t)} \) is the cost of VMs in the on-demand scheme.

We take the RTT as the QoS factor of multimedia applications. In time slot \( t \), the RTT of group-\( r \) users for type-\( j \) application can be formulated as \( T_{rj}^{RTT(t)} = T_{rj}^{fud(t)} + T_{rj}^{res(t)} + T_{rj}^{bad(t)} \), where \( T_{rj}^{res(t)} \) is the service response time of type-\( j \) application. As presented in Section III, the request arrivals for type-\( j \) application follow a Poisson Process with mean arrival rate of \( \lambda_{ij}^{(t)} \). For type-\( j \) application, \( (K_{ij}^{res(t)} + K_{ij}^{dem(t)}) \) class-\( i \) VMs are allocated for serving requests. To balance the workload among different classes virtual clusters, we use the normalized service rate ratio [19] as the scheduling probability, i.e., one request for type-\( j \) application is scheduled to class-\( i \) virtual cluster with the probability \( \omega_{ij}^{(t)} = \frac{(K_{ij}^{res(t)} + K_{ij}^{dem(t)}) \mu_{ij}}{\sum_{i=1}^{N} (K_{ij}^{res(t)} + K_{ij}^{dem(t)}) \mu_{ij}} \). According to the decomposition property of Poisson distribution [18], the arrivals of scheduled requests for type-\( j \) application at class-\( i \) virtual cluster also follow a Poisson Process with mean arrival rate of \( \omega_{ij}^{(t)} \lambda_{ij}^{(t)} \). In addition, the service time of the class-\( i \) VMs is assumed to be exponentially distributed with mean service time of \( \frac{1}{(K_{ij}^{res(t)} + K_{ij}^{dem(t)}) \mu_{ij}} \). Thus, the service process of class-\( i \) VMs in type-\( j \) application can be modelled as an \( \text{M/M/1} \) queuing system [18]. To make the queue stable, \( \omega_{ij}^{(t)} \lambda_{ij}^{(t)} < (K_{ij}^{res(t)} + K_{ij}^{dem(t)}) \mu_{ij} \) should be satisfied. The service response time of type-\( j \) application at class-\( i \) VMs can be formulated as \( T_{ij}^{resp(t)} = \frac{1}{(K_{ij}^{res(t)} + K_{ij}^{dem(t)}) \mu_{ij} - \omega_{ij}^{(t)} \lambda_{ij}^{(t)}} \), and thus the mean service response time of type-\( j \) application is given by \( T_{ij}^{resp(t)} = \frac{\sum_{r=1}^{R} \omega_{ij}^{(t)} \mu_{ij} r_{ij}^{resp(t)}}{\omega_{ij}^{(t)} \lambda_{ij}^{(t)} + \sum_{r=1}^{R} \omega_{ij}^{(t)} \mu_{ij} r_{ij}^{resp(t)}} \).

The total number of allocated class-\( i \) VMs in the reservation plan should be no more than the number of initially reserved class-\( i \) VMs. The constraints can be formulated as \( \sum_{j=1}^{M} K_{ij}^{res(t)} \leq K_{ini}^{i} \), \( (\forall i = 1, 2, \ldots, N) \).

Based on the above analysis, we can formulate the VM allocation problem for single-site cloud as follows.

\[
\begin{align*}
\text{Minimize} \quad & \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij}^{res} K_{ij}^{res(t)} + \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij}^{dem} K_{ij}^{dem(t)} \\
\text{subject to} \quad & T_{rj}^{fud(t)} + \sum_{i=1}^{N} \sum_{j=1}^{M} \omega_{ij}^{(t)} \mu_{ij} r_{ij}^{resp(t)} \\
& + T_{rj}^{bad(t)} \leq T_{rj}^{resp(t)}, \quad \forall r = 1, \ldots, R, \\
& \forall j = 1, \ldots, M, \\
& \omega_{ij}^{(t)} \lambda_{ij}^{(t)} < (K_{ij}^{res(t)} + K_{ij}^{dem(t)}) \mu_{ij}, \quad \forall i = 1, \ldots, N, \quad \forall j = 1, \ldots, M, \\
& \sum_{j=1}^{M} K_{ij}^{res(t)} \leq K_{ini}^{i}, \quad \forall i = 1, \ldots, N,
\end{align*}
\]

where \( T_{rj}^{resp(t)} \) is the upper bound of the RTT for type-\( j \) application.

The optimal VM allocation problem (1) is an integer programming, which is known to be NP-hard [20]. The problem can be solved by the branch-and-bound method [20]. However, the time complexity of the branch-and-bound cannot satisfy the efficiency demand of the practical VM allocation. Therefore, we propose the greedy algorithm to allocate VMs for single-site cloud, which is presented in Algorithm 1.

### V. OPTIMAL VM ALLOCATION SCHEME FOR MULTI-SITE CLOUD

In this section, we investigate the VM allocation problem for multi-site cloud, in which the application providers can deploy services on geographically distributed data centers to reduce the transmission delay between users and data centers. For instance, Amazon EC2 [2] provides seven data centers all over the world. In different data centers, the reservation and on-demand price rates are different. We assume that users choose the closest data center for service to save the transmission delay.

Suppose that there are \( L \) data centers at different locations, and \( D = \{D_{1}, D_{2}, \ldots, D_{L}\} \) is the set of data centers. Let \( p_{ij}^{res} \) and \( p_{ij}^{dem} \) denote the reservation and on-demand price rates for class-\( i \) VM instance at data center \( l \) (\( \forall l = 1, 2, \ldots, L \)), respectively. Since users receive service from the closest data center, we denote the user set for data center \( l \) as \( R_{l} \), in which the users in group-\( r \) (\( r \in R_{l} \)) receive service from data
Algorithm 1 Greedy Algorithm for VM Allocation in Single-site Cloud

1: Initialization: Compute $q_{ij}^{\text{res}} = \frac{p_{ij}^{\text{res}}}{p_{ij}}$ and $q_{ij}^{\text{dem}} = \frac{p_{ij}^{\text{dem}}}{p_{ij}}$, which are the cost-performance ratios by using reserved or on-demand class-\(i\) VMs for processing type-\(j\) application, and set $Q = \{q_{ij}^{\text{res}}, q_{ij}^{\text{dem}}, \ldots, q_{LM}^{\text{res}}, q_{LM}^{\text{dem}}\}$. Set $\lambda_{\text{rem}}(t) = \{\lambda_1^{\text{rem}}(t), \ldots, \lambda_M^{\text{rem}}(t)\}$, in which $\lambda_j^{\text{rem}}(t) = \sum_{r=1}^R \lambda_j^r$, represents the unprocessed requests for type-\(j\) application in the decision slot $t$.

2: Sort set $Q$ in ascending order.

3: repeat
4: Select the smallest $q_{ij}^v$ ($\forall v = 1, \ldots, N, \forall j = 1, \ldots, M$) from on-demand VMs (i.e. $v = \text{dem}$) then
5: Allocate $K_{ij}^{\text{res}(t)}$ class-\(i\) reserved VMs to process requests $\lambda_j^f(t) \leq \lambda_j^{\text{rem}}(t)$ as long as the requirements $T_{r}^{\text{fwd}(t)} + \frac{1}{K_{ij}^{\text{res}(t)} \mu_j - \lambda_j^f} + T_{r}^{\text{bwd}(t)} \leq \tau_j$ and $K_{ij}^{\text{res}(t)} \leq K_{ij}^{\text{ini}}$ are both satisfied. Update $K_{ij}^{\text{ini}} = K_{ij}^{\text{ini}} - K_{ij}^{\text{res}(t)}$, $\lambda_j^{\text{rem}}(t) = \lambda_j^{\text{rem}}(t) - \lambda_j^f$,
6: else if $q_{ij}^v$ from on-demand VMs (i.e. $v = \text{dem}$) then
7: Allocate $K_{ij}^{\text{dem}(t)}$ class-\(i\) on-demand VMs to process all remaining type-\(j\) application requests $\lambda_j^{\text{rem}}(t)$ until the requirement $T_{r}^{\text{fwd}(t)} + \frac{1}{K_{ij}^{\text{dem}(t)} \mu_j - \lambda_j^{\text{rem}}(t)} + T_{r}^{\text{bwd}(t)} \leq \tau_j$ is satisfied. Update $\lambda_j^{\text{rem}}(t) = 0$,
8: end if
9: until all requests are processed.
10: Calculate the total cost $C_{\text{sim}} = \sum_{i=1}^L \sum_{j=1}^M p_{ij}^{\text{res}} K_{ij}^{\text{res}(t)} + \sum_{i=1}^L \sum_{j=1}^M p_{ij}^{\text{dem}} K_{ij}^{\text{dem}(t)}$.

The optimal VM allocation for multi-site cloud can be formulated as follows.

Minimize $\{K_{ij}^{\text{res}(t)}, K_{ij}^{\text{dem}(t)}\}$ subject to

$$
L \sum_{i=1}^L \sum_{j=1}^M p_{ij}^{\text{res}} K_{ij}^{\text{res}(t)} + \sum_{i=1}^L \sum_{j=1}^M p_{ij}^{\text{dem}} K_{ij}^{\text{dem}(t)}$$

$$
T_{r}^{\text{fwd}(t)} + \frac{1}{K_{ij}^{\text{res}(t)} \mu_j - \lambda_j^{\text{rem}(t)}} + T_{r}^{\text{bwd}(t)} \leq \tau_j,
\forall j = 1, \ldots, M, \forall i = 1, \ldots, L, \forall r \in R_i
$$

$$
\omega_{ij}^{(t)} \sum_{r \in R_i} \lambda_j^r < (K_{ij}^{\text{res}(t)} + K_{ij}^{\text{dem}(t)}) \mu_j,
\forall i = 1, \ldots, L,
$$

$$
\omega_{ij}^{(t)} \sum_{r \in R_i} \lambda_j^r + (K_{ij}^{\text{res}(t)} + K_{ij}^{\text{dem}(t)}) \mu_j,
\forall i = 1, \ldots, L,
$$

where $\tau_j$ is the upper bound of the RTT for type-\(j\) application.

The optimal VM allocation problem (2) is an integer programming, which is known to be NP-hard [20]. The problem can be solved by the branch-and-bound method [20]. However, the multimedia application providers require rapid and efficient VM allocation scheme, which can quickly adapt to the time-varying workload. Therefore, we propose the greedy algorithm for VM allocation in multi-site cloud, which can be presented in Algorithm 2.

Algorithm 2 Greedy Algorithm for VM Allocation in Multi-site Cloud

1: Sort the data centers set $D = \{D_1, D_2, \ldots, D_L\}$ in descending order based on the number of requests.
2: repeat
3: Select the data center with the most requests.
4: Allocate VMs for the selected data center using the Greedy Algorithm for VM Allocation in Single-site Cloud.
5: Remove the selected data center from set $D$.
6: until all requests in all data centers are processed.
VI. SIMULATIONS

A. Simulations for Single-site Cloud

1) Parameter Settings: We perform simulations to evaluate the proposed optimal VM allocation scheme for single-site cloud. Amazon EC2 [2] is the Amazon’s cloud computing platform allowing application providers to rent VMs for their applications. To make our evaluation convincible, we employ the price rates and VM configurations of Amazon EC2. We select the Amazon’s data center in Ireland as the single data center in our simulation. Four classes of VMs are provided. The reservation and on-demand price rates of the four classes VMs are \( p^{res} = \{0.063\$/h, 0.25\$/h, 0.50\$/h, 0.75\$/h\} \) and \( p^{dem} = \{0.095\$/h, 0.38\$/h, 0.76\$/h, 1.14\$/h\}, \) respectively. The numbers of initially reserved VMs in each class are \( K^{ini} = \{240, 120, 30, 18\}. \) The application provider provides five multimedia applications. The service rates of each class VM for processing each type application are randomly distributed. The RTT requirements of the multimedia applications are \( \tau = \{0.5s, 0.6s, 0.7s, 0.8s, 0.9s\}. \) We select six groups of users at different locations in our simulation. The forward and backward transmission delays are randomly distributed. The mean arrival rates of users’ requests vary with time. The time slot for VM allocation is set as 1 hour.

2) Simulation Results: We first compare the resource cost between the proposed optimal VM allocation scheme, in which the VMs are allocated optimally for each type application by solving the optimization problem (1), and the proposed greedy algorithm for VM allocation in single-site cloud. The optimal VM allocation scheme is global optimal benchmark but not practical, while the greedy algorithm is sub-optimal but lightweight and practical.

When the mean request arrival rate varies from 5000 requests/s to 15000 requests/s, the comparison of the resource cost between the proposed optimal VM allocation scheme and the proposed greedy algorithm is shown in Fig. 2(a). From Fig. 2(a), we can see that the proposed optimal VM allocation scheme can achieve lower resource cost compared to the greedy algorithm under the same request arrival rate. The greedy algorithm only considers the current best choice but fails to make a global inspection, while the optimal VM allocation scheme searches the whole feasible region to achieve the global optimal solution. Thus, the greedy algorithm achieves the sub-optimal solution with lightweight. Fig. 2(b) shows the comparison of resource cost in each application when the total request arrival rate is 15000 requests/s.

We record the service process of the data center in a 9-hour period. During the period, the mean request arrival rate is varied as in Fig. 3(a). The corresponding resource cost between the proposed optimal VM allocation scheme and the greedy algorithm is shown in Fig. 3(b). From Fig. 3(b), we can find that the maximal difference of the resource cost between the optimal VM allocation scheme and the greedy algorithm is 4.75 dollars at the 6th hour.

B. Simulations for Multi-site Cloud

1) Parameter Settings: In this subsection, we perform simulations to evaluate the proposed optimal VM allocation scheme for multi-site cloud. We employ three geographically distributed Amazon data centers, in which D1 is located at Asia Pacific region Tokyo, D2 is located at EU region Ireland which is also used in Section VI-A1, and D3 is located at US region East Virginia. The price rates in D2 are already provided in Section VI-A1. The reservation and on-demand price rates in D2 are \( p^{res} = \{0.069\$/h, 0.275\$/h, 0.55\$/h, 0.83\$/h\}, \) \( p^{dem} = \{0.108\$/h, 0.40\$/h, 0.80\$/h, 1.20\$/h\}, \) \( p^{res} = \{0.055\$/h, 0.25\$/h, 0.45\$/h, 0.57\$/h\}, \) \( p^{dem} = \{0.085\$/h, 0.34\$/h, 0.68\$/h, 1.00\$/h\}. \) The numbers of initially reserved VMs in each data center are the same, \( K^{ini} = \{80, 40, 10, 6\}. \) The total initially reserved VMs are the same to the single-site cloud case. Since users choose the closest data center for service, we set that D1 is the closest data center for group-1 and group-2 users, D2 is the closest data center for group-3 and group-4 users, and D3 is the closest data center for group-5 and group-6 users. The forward and backward transmission delays from users to the closest data center are randomly distributed. The other parameters including the time slot and the RTT requirements are the same as the settings in Section VI-A1.

2) Simulation Results: We compare the resource cost between the proposed optimal VM allocation scheme for multi-site cloud, in which the allocated VMs are determined by solving the optimization problem (2), and the proposed greedy
Fig. 4. Comparison results for multi-site cloud. (a) Comparison of resource cost between the optimal VM allocation scheme and the greedy algorithm. (b) Comparison of resource cost between the optimal VM allocation schemes for single-site cloud and for multi-site cloud.

Fig. 5. Simulation results for multi-site cloud in a 9-hour period. (a) Comparison of resource cost between the optimal VM allocation scheme and the greedy algorithm in the period. (b) Comparison of resource cost in each data center at the 6th hour.

algorithm for VM allocation in multi-site cloud. The optimal VM allocation scheme provides the global optimal benchmark but not practical, while the greedy algorithm is sub-optimal but lightweight and practical.

Fig. 4(a) shows the comparison of resource cost when the request arrival rate increases from 5000 requests/s to 15000 requests/s. From Fig. 4(a), we can see that the proposed optimal VM allocation scheme can provision service at less resource cost than the greedy algorithm. We also compare the cost of the optimal VM allocation scheme for single-site cloud and that for multi-site cloud, which is shown in Fig. 4(b). From Fig. 4(b), we can find that under the same request arrival rate, the resource cost in the multi-site cloud is lower than that in the single-site cloud. In the multi-site cloud, users can request application service from the closest data center, which greatly reduces the transmission delay, thus the resource demands are decreased, leading to the less resource cost than that in the single-site cloud.

We also perform evaluation in a 9-hour period. The request arrival rate during this period is the same as that in Fig. 3(a). The corresponding resource costs of the optimal VM allocation scheme and the greedy algorithm are shown in Fig. 5(a). From Fig. 5(a), we can see that the resource cost in the greedy algorithm is close to that in the optimal VM allocation scheme. The minimal difference of resource cost between these two schemes is 0 dollars at the 1st hour, while the maximal difference is 5.6 dollars at the 6th hour. Fig. 5(b) shows the comparison of resource cost in each data center at the 6th hour.

Compared with the greedy algorithm, the proposed optimal VM allocation scheme achieves lower resource cost at every data center.

VII. CONCLUSIONS

In this paper, we investigate the optimal VM allocation problem for multimedia application providers to provide services at the minimal resource cost. Based on the SaaS architecture and the VM allocation model, we propose the optimal VM allocation schemes for single-site cloud and multi-site cloud, respectively. In each case, we formulate and solve the VM allocation problem to minimize the resource cost under the RTT constraints. Moreover, we propose the greedy algorithm in each case to efficiently allocate VMs for the practical usage. The simulation results demonstrate that the proposed optimal VM allocation schemes can effectively achieve the minimal resource cost for multimedia application providers.

REFERENCES