Auction Mechanism for Dynamic Bandwidth Allocation in Multi-Tenant Edge Computing

Tra Huong Thi Le*, Nguyen H. Tran†, Tuan LeAnh*, Thant Zin Oo*, Kitae Kim*, Shaolei Ren‡,
Choong Seon Hong*

*Department of Computer Science and Engineering, Kyung Hee University, South Korea
†School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia
‡Department of Electrical and Computer Engineering, University of California at Riverside, Riverside, CA 92505 USA

Abstract—Inspired by the shared infrastructure of colocation data centers and the growth of Mobile Edge Computing (MEC), colocation MEC businesses have thrived to offer an economical and low latency solution for MEC based tenants. In colocation MEC, a colocation service provider leases spaces at the base stations (BSs) to tenants for housing their servers. Although the tenants fully manage their own servers, they still need to purchase bandwidth from the network provider to via bandwidth purchasing market to serve their users. As a result, tenants should consider the trade-off between delay and energy consumption, which are related to bandwidth and server computing speed, respectively. In this paper, an auction framework is used to model the interaction between the network provider and tenants in the bandwidth market. The network provider determines the bandwidth amount and corresponding payment by tenants based on their submitted bids, in which the valuation of the bandwidth amount in each bid is based on the coupling between bandwidth and server computing speed. The networking provider also seeks winners in the social welfare maximization problem, which is NP-hard. To solve this problem, we propose a solution based on the randomized auction mechanism, which is proved to be approximately truthful, individually rational, and computationally efficient. Simulation results verify the effectiveness of our proposed randomized auction scheme in bandwidth trading.

Index Terms—Mobile Edge Computing, Colocation, Bandwidth Trading, Auction.

I. INTRODUCTION

In recent years, the emergence of Mobile Edge Computing (MEC) is considered as the evolution of the cloud computing where the computational power is distributedly deployed in small data centers at the edge network [1]–[5]. With proximity access, MEC can reduce access delay, allowing for a wide range of new time-critical applications and services, such as real-time online gaming, virtual reality, and ultra-high-definition video streaming. The shared infrastructure of a colocation data center motivated us to envision colocation MEC in which the colocation service provider offers leased micro data centers at base stations (BSs) and houses its tenants’ servers [6]. Colocation MEC enables tenants to fully control their own servers without the additional burden of non-IT power/cooling infrastructure management. Toward this trend, the colocation service provider has a design requirement for tenant-driven micro data centers and remote infrastructure management software, which can provide many telemetry data for tenants during application workload deployment or movement [7]. One example of colocation service provider is Vapor IO, which deploys cylindrical data center enclosures called Vapor Chambers at base stations in partnership with wireless tower companies [8]. According to a recent survey by Uptime Institute [9], 75% of respondents are willing to house their physical servers at colocation MEC sites.

Colocation MEC is advantageous in many respects. First, tenants such as companies focusing on high-tech applications (i.e., drone business) do not need to invest in new non-IT physical infrastructures to explore a new market. Second, the tenants have full control of their own physical servers, which results in no latency interference from other tenants. Third, because the tenants do not need to manage the non-IT power/cooling infrastructure, they have better scalability and spend lower capital/operational costs compared to self-built data centers. Fourth, hyperscale cloud providers also benefit from colocation MEC, as they can expand their businesses without building or managing non-IT infrastructures at every edge location. Fifth, colocation MEC is also a revenue opportunity for network carriers such as Verizon and AT&T [7], [10], as the tenants require bandwidth to transmit the computing results to their users. The more tenants participate in colocation MEC, the more revenue the network provider receives. Furthermore, distribution in colocation MEC has higher Quality of Service (QoS) in terms of latency (delay) with decreased cost due to statistical multiplexing gain and resource efficiency.

However, the deployment of colocation MEC faces several challenges. One key challenge is bandwidth allocation among tenants, as the bandwidth of the network provider is limited. Thus, the network provider is motivated to develop bandwidth sharing among tenants to improve bandwidth efficiency and achieve maximal social welfare. Furthermore, because the
current static bandwidth allocation is ineffective for coping with dynamic traffic demand, the network provider requires a dynamic bandwidth sharing scheme for allocating the bandwidth to tenants.

Another challenge of colocation MEC is the problem of asymmetric information of the network provider and tenants. In this paper, we take the viewpoint that network provider allocates bandwidth to self-interested tenants. For tenants, low service latency must be guaranteed to attract more users to offload their computing tasks to MEC servers. Latency for the mobile service consists of two elements: computation latency, which depends on the CPU-cycle frequency, and communication latency, which depends on the data transmission rate [11], [12]. While the bandwidth allocation scheme determines the data rate, the CPU-cycle frequency can be adjusted, which also determines the energy consumption of MEC servers. The computation time can be reduced, in particular by increasing the CPU-cycle frequency; however, doing so leads to higher energy consumption. Therefore, the optimal bandwidth amount that one tenant rents from the network provider should be based on the trade-off between latency and energy consumption. However, the CPU-cycle frequency is a private decision made by the individual tenant and is unknown to the network provider. Tenants may be self-interested and provide an incorrect valuation of their required bandwidth to receive more bandwidth. Therefore, the bandwidth allocation scheme of the network provider must to incentivize tenants to reveal truthful information.

To address the above challenges, a feasible solution based on the auction mechanism is proposed, in which bidders (tenants) submit their bids to an auctioneer (the network provider) to buy bandwidth. Then, the network provider decides which bid is accepted so that both the network provider and tenants are satisfied with the incentive and the social welfare is maximized.

There are three properties characterize an auction. Those are: social efficiency, which is defined as the sum of utilities of all players, individual rationality, which ensures that the utility of each individual player is non-negative, and truthfulness, which incentivizes bidders to reveal the true value in submitted bids. These properties guarantee that bandwidth is shared efficiently and fairly among the tenants in the network. Of these properties, truthfulness is the most important [13]. The Vickrey-Clarke-Groves (VCG) auction ensures truthfulness and social efficiency if a social welfare optimization problem can be solved optimally [14], [15]. However, the optimization problem in this study is NP-hard. Furthermore, the truthfulness of the VCG mechanism can be compromised if the optimization problem is solved approximately [16]. In this paper, we propose a mechanism for purchasing bandwidth in colocation MEC. The proposed mechanism is based on the randomized auction [17], [18] and is proved to be approximately truthful, individually rational, and computationally efficient.

The main contributions of this paper are as follows:

- We consider a colocation MEC system, which consists of one network provider and multiple tenants that house their servers at rack spaces leased from a colocation service provider. We model the bandwidth trading market between the network provider and tenants as an auction problem. In each bid submission to the network provider, the tenant calculates a valuation of its required bandwidth coupled with its servers’ computing speed. Each bid is affected by the tenant’s desire to minimize delay and energy consumption.

- We demonstrate that the winner determination problem in the auction problem is NP-hard, and we propose a solution based on a randomized auction mechanism. The solution mechanism has two components; one for winner selection and the other for payment determination. The outputs of both components are involved in the output of fractional VCG mechanism and the upper bound of the integrality gap. A greedy approximation algorithm is proposed to obtain the upper bound of the integrality gap and is used to decompose the winner selection output of the fractional VCG mechanism into a set of feasible integer solutions of the original winner determination problem. We also demonstrate that the proposed scheme guarantees social efficiency, individual rationality, approximate truthfulness, and computational efficiency.

- We run extensive experiments to verify the performance of the proposed mechanism in bandwidth purchasing for colocation MEC. Compared to the existing randomized auction, the proposed mechanism achieves 2% higher average social welfare and 3.1% higher utilization for a tenant population of 100.

The remainder of this paper is organized as follows. In Section II, we briefly review some related works, while in Section III, we introduced the system model. We describe the problem formulation in Section IV, and present the randomized auction-based bandwidth purchasing in Section V. We present simulation results in Section VII, and provide conclusion in Section VIII.

II. RELATED WORK

Most research on the MEC focuses on the offloading strategy and resource allocation in a homogeneous system [19], [21]–[27]. The aim is to minimize system energy consumption and/or the time taken to complete computing tasks. Pham et al. [19] jointly optimized the offloading strategy, transmission power, and the servers’ resource allocation to minimize the computational overhead in a MEC system. Zhang et al. [21] studied minimizing the energy consumption in computing tasks and transmitting files considering a 5G heterogeneous network environment. The authors of [22] examined how mobile terminals compete over limited communication resources in HetNets and limited computation resources in a MEC server. Authors in [25] investigated the offloading decision problem of IoT mobile devices in ultra-dense IoT networks. The work in [25] took into account the diversity of MDs’ computation tasks, the randomness of task arrival at the edge servers, and the dynamic variation of the computing resources at the edge servers. In [25], the system model was presented in terms of the MDs’ energy consumption and the computation tasks’ processing delay, and the problem aimed to...
minimize the overall computation overhead while the wireless channel constraints were satisfied. The proposed solution to solve the problem in [25] is the two-tier game-theoretical greedy approximation offloading scheme, which provides high computational efficiency. The work in [26] considered cloudlet resources’ optimization problem to maximize system scalability and make Service Delay (which encompasses Transmission Delay, Processing Delay and Backhaul Delay) as low as possible. The authors in [26] utilized an analytical model of the Service Delay and Particle Swarm Optimization algorithm to reconfigure the system, and then they utilized Transmission Power Control and Virtual Machine (VM) migration to realize and enable the configuration. However, they did not consider the cost of VM migration. Due to the significant growth of the number of devices, servers and applications, the MEC has issue with problem dimensionality, number of parameters and solution executability. One solution is the utilization of machine learning (ML). In [27], the authors provided a comprehensive survey and helpful guidance on the use of ML in MEC systems. In [23], task offloading and resource allocation in proximity clouds were considered. The authors solved the maximization problem in quality-of-experience between offloading and local execution. The work in [24] considered the offloading decision and virtual machine (VM) level allocation problems in multi-hop multi-user blockchain empowered MEC to maximize long-term offloading performance. They proposed the model-free deep reinforcement learning algorithm to adapt to highly dynamic environments. The proposed algorithm takes advantage of the genetic algorithms into deep reinforcement learning during the exploration process in order to increase the convergence speed. However, the work sets the maximum achievable rate for downloading/uploading as 250 kbps. The work does not consider the bandwidth allocation. Moreover, the set of VM levels, VM levels’ prices, CPU cycles are predefined.

Other works have studied pricing in a MEC system [28], [29]. The authors of [28] considered a single-cloud multi-user MEC system and modeled the problem of computation capacity pricing as a single-leader-multi-user Stackelberg game taking into account the resource requirements of each mobile user in the system. The authors of [29] proposed a hierarchical model in which the level of cloudlets is based on the principle of an LTE-advanced backhaul network. The service provider uses a two-timescale resource allocation scheme in which computing resources are allocated on a longer timescale, and bandwidth for communication is allocated on a shorter timescale. The authors formulated the problem of profit maximization for the service provider. Then, they designed a heuristic algorithm to solve binary linear programming. However, they did not prove that the proposed algorithm was truthful, which is an essential property of the auction game. In [30], a heterogeneous multi-layer mobile edge computing (HetMEC) was proposed. The computing task can be divided into several parts to be processed at mobile devices, MEC servers, and cloud centers. To support a low latency service and capture the time variation of system status, a reinforcement learning (RL)-based framework for HetMEC was proposed. The authors in [30] also studied pricing, which is one of the extensive research issues based on the proposed framework. To incentive mobile devices and MEC servers to carry out the computing tasks, the cost of the computing/transmission resources of each device should be involved in the action space and the price should be a component of the environment state. However, the author in [30] did not mention designing an incentive mechanism in detail. This study is the first to examine the bandwidth sharing problem for a heterogeneous colocation MEC system in which many tenants share spaces in the tower of a network operator for housing their computing servers and share the total bandwidth of the network operator. In this paper, we also prove that the proposed auction-based bandwidth purchasing scheme is approximately truthful.

Auction-based approaches have attracted much research interest for efficient spectrum trading in cognitive radio networks, device-to-device (D2D) networks, and so on. The authors of [31] introduced a double-sided bandwidth auction game for cognitive D2D communication. In the auction, the D2D transmitter bids a demand price-bandwidth curve, and the service providers offer a price-bandwidth curve. The objective is to minimize the payout, which is dependent on the revenue of the D2D users per unit of the transmission rate and the spectral efficiency of transmission. In [32], the channel allocation problem among bidders was considered in two separate cases: single-channel and bundles of channels. A VCG mechanism [33] was proposed to solve the problem; however, the author could not prove the truthfulness of the proposed methods. Zhu et al. [34] designed core-selecting auctions, which prevent collusion and shill bidding often seen in VCG and improve seller revenue. However, the authors could not guarantee absolute truthfulness, although their proposal prevented deviations from truthfulness.

The authors of [35] proposed a general truthful double spectrum auction framework called TRUST. Buyers were divided into multiple independent sets or non-interfering groups. Each group was treated as one virtual buyer. Then, McAfee [36] was applied. All members of the group shared the same price. TRUST can achieve truthfulness and other auction properties while improving spectrum utilization; however, its efficiency is low, and its fairness is poor. In these existing studies, the valuation and utility in auctions are often selected in an ad hoc manner or based on the unique property of spectrum resources: interference. However, tenants in a MEC system should focus on the delay constraint and energy consumption for economic benefit. In this paper, due to the relationship between the delay cost and energy consumption of MEC systems, we consider the valuation of bandwidth trading of each tenant as a monotonically decreasing function of the weighted cost (delay and energy consumption). Another difference between this study and previous studies is that this study uses a randomized auction in which different tenants are allocated randomly with different bandwidth amounts in different time slots. The work in [37] considered the optimal computing allocation problem in edge cloud assisted Internet of Things in blockchain. The problem was formulated as the double-sided auction in which the broker is the auction controller and guarantees sellers and buyers of computing resources. However, there are many iterations among the broker, sellers, and buyers to determine...
how much computing resource is traded and the corresponding price to guarantee the truthfulness. Therefore, there are several communication overheads in the system. Moreover, different from our work, the allocated resources to buyers in [37] are decided by the broker based on the initial prices submitted by the buyers and sellers and the underlying optimization problems are offline convex. The paper [38] considered the problem of motivating the mobile devices (MD) to process tasks for IoT devices. The authors proposed an online reward-optimal auction (RoA) to maximize the long term sum of rewards of the edge system without any prior knowledge of the energy harvested process, task arrival and channel condition. RoA is designed based on Lyapunov optimization and VCG auction. However, different from our work, this paper did not account for the delay requirement in the value function applied in MDs’ utility functions.

III. SYSTEM MODEL

As illustrated in Fig. 1, we consider a network provider such as T-mobile, that operates an edge data center and provides bandwidth to its $K$ tenants that provide time-critical applications. Table I presents the notations used in this paper. According to the virtualization-based resource sharing in [39], the network provider slices the entire bandwidth at the considered base station (BS) and allocates its to tenants so that they can transmit and receive the computing tasks of their users associated with the BS. It is assumed that at the BS, the network provider has a bandwidth of $Y$ Hz to lease. In our model, the network provider allocates the bandwidth of the BS to tenants in a hybrid manner (both statically and dynamically). In static bandwidth allocation, each tenant $k$ reserves bandwidth $y_k$ for a long duration and makes a corresponding payment in advance. This is performed by making a long-term prediction of tenant traffic demand from its traffic usage history. Determining the reserved demand and corresponding payment is assumed to be predefined. In dynamic bandwidth allocation, the network provider leases available bandwidth to tenants on a short timescale to reflect the time-varying traffic demand. The available dynamic bandwidth for the network provider (i.e., the remaining bandwidth after static allocation) is given as follows:

$$W = Y - \sum_{k \in K} y_k$$  \hspace{1cm} (1)

Then the real-time demand of tenant $k$ can be given by

$$z_k = \max (d_k - y_k, 0),$$

where $d_k$ is the bandwidth required by tenant $k$, which is estimated based on the current information. Bandwidth trading between the network provider and tenants can be formulated as an auction game. In the auction, the network provider, which acts as an auctioneer leases available dynamic bandwidth to the tenants, which act as bidders. Each tenant $k \in K$ can submit $J_k \geq 1$ bids at each BS. Submitting multiple bids is a general case and includes "single bid" as a special case. Each bid from tenant $k$ consists of the desired bandwidth amount $n_{kj}$ such that $0 \leq n_{kj} \leq z_k$, along with the corresponding valuation $b_{kj}$. Let $t_{kj} = \{n_{kj}, b_{kj}\}$ denote the $j$th bid submitted by tenant $k$ submitted and $T_k = \{t_{kj}\}$ denote the set of $J_k$ biddings of tenant $k$. It should be noted that the true valuation $V_{kj}$ and the true demanded dynamic resource vector $z_k$ are private information belonging to tenant $k$ and are unknown by the network provider and other tenants. Thus, $b_{kj}$ may differ from $V_{kj}$. However, in a truthful auction, a bidder must reveal the true valuation $b_{kj} = V_{kj}$ to obtain the highest utility. The means of determining the true valuation $V_{kj}$ is explained later in section IV-A. Let $x_{kj}$ be a variable indicating whether tenant $k$’s $j$th bid wins.

$$x_{kj} = \begin{cases} 1, & \text{if tenant } k\text{'s } j\text{th bid wins} \\ 0, & \text{otherwise.} \end{cases}$$  \hspace{1cm} (2)

We adopt the XOR bidding language, in which a tenant can win at most one bid although it can submit multiple bids. Thus, the first constraint for bandwidth allocation is given as follows:

$$\sum_{j=1}^{J_k} x_{kj} \leq 1, \quad \forall k \in K,$$  \hspace{1cm} (3)

We also assume that the total allocated bandwidth does not exceed the bandwidth capacity of the network provider at the BS, as follows:

$$\sum_{k=1}^{K} \sum_{j=1}^{J_k} n_{kj} x_{kj} \leq W.$$  \hspace{1cm} (4)

The utility of tenant $k$’s $j$th bid is given as

$$u_{kj} = \begin{cases} V_{kj} - p_{kj}, & \text{if } x_{kj} = 1, \\ 0, & \text{otherwise.} \end{cases}$$

In this paper, the tenants’ workload arrival rates are provided externally. In practice, a tenant’s workload arrival rate in edge computing is determined by its users’ computation offloading decisions, which can be partial offloading (i.e., an application is partially executed by a user’s device and partially by edge servers) or full offloading (i.e., an application is entirely run on edge servers). Because in this study we focus on tenants’ resource management and bandwidth acquisition decisions, explicitly modeling users’ offloading decisions is beyond the scope of this paper.
TABLE I: Table of notation

<table>
<thead>
<tr>
<th>Description</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenant index</td>
<td>k</td>
</tr>
<tr>
<td>Bid index</td>
<td>j</td>
</tr>
<tr>
<td>Available dynamic bandwidth at</td>
<td>W</td>
</tr>
<tr>
<td>Bidding bandwidth amount of</td>
<td>$n_{kj}$</td>
</tr>
<tr>
<td>jth bid of tenant k</td>
<td>$b_{kj}$</td>
</tr>
<tr>
<td>Valuation of jth bid of tenant k</td>
<td>$V_{kj}$</td>
</tr>
<tr>
<td>True valuation of jth bid of</td>
<td>$p_{kj}$</td>
</tr>
<tr>
<td>tenant k</td>
<td></td>
</tr>
<tr>
<td>Transmission delay of tenant k</td>
<td>$D_{kj}^p$</td>
</tr>
<tr>
<td>Processing delay of tenant k</td>
<td>$D_{kj}^p$</td>
</tr>
<tr>
<td>Delay cost of tenant k</td>
<td>$C_{kj}$</td>
</tr>
<tr>
<td>Energy cost of tenant k</td>
<td>$E_{kj}$</td>
</tr>
<tr>
<td>Fractional VCG based payment of</td>
<td>$P_{kj}$</td>
</tr>
<tr>
<td>the jth bid of tenant k</td>
<td></td>
</tr>
<tr>
<td>Integrality gap</td>
<td>$\alpha$</td>
</tr>
</tbody>
</table>

We denote the arrival rate of a task of tenant $k$ by $\lambda_k$. In addition, we assume that the task requested by the users of tenant $k$ has a data size following an exponential distribution of mean $L_k$ and requires a CPU cycle of $C_k$. If tenant $k$’s $j$th bid wins the auction or the bandwidth amount allocated to tenant $k$ is $n_{kj}$, the achievable downlink rate of a typical user of tenant $k$ is as follows:

$$R_{kj} = (n_{kj} + y_k) \log_2(1 + \Gamma_k).$$

(5)

where $\Gamma_k = \frac{p}{N_0(n_{kj} + y_k)}$ is the signal-to-noise-ratio and $N_0$ is the power density of the noise. Tenants must transfer the computing results to their users.

There are two factors affecting the delay of transferring the computing results to the tenants’ users: the processing delay and the transmission delay. We model the transmission delay of tenant $k$ as follows:

$$D_{kj}^t = \frac{L_k\lambda_k}{R_{kj}}.$$  

(6)

We denote $s_{kj}$ as the corresponding CPU cycle of tenant $k$ when the purchasing bandwidth is $n_{kj}$. The processing delay of tenant $k$ is as follows:

$$D_{kj}^p = \frac{C_k\lambda_k}{s_{kj}}.$$  

(7)

The total delay of a user-requested service of tenant $k$ is the sum of the processing delay at the servers and the transmission delay, and is given as follows:

$$D_{kj} = D_{kj}^p + D_{kj}^t.$$  

(8)

Each tenant $k$ has a delay constraint. If the total delay of each user of tenant $k$ exceeds this delay constraint, tenants $k$ incur a delay cost. In addition, we assume that the tenant determines its server resource provisioning and bandwidth acquisition according to the single time-scale. For example, given the bandwidth, the tenant can increase its CPU speed very quickly, which yields a better delay performance at the expense of increased energy consumption. Therefore, together with the delay cost, each tenant incurs server energy consumption cost. Thus, we present two types of costs associated with tenants: a delay cost and an energy cost.

Delay cost is defined as follows:

$$C_{kj} = \begin{cases} \beta_k(D_{kj} - D_{th}^k), & \text{if } D_{kj} \geq D_{th}^k; \\ 0, & \text{otherwise}, \end{cases}$$  

(9)

where $D_{th}^k$ is the delay constraint of tenant $k$ [40]. We consider a soft delay threshold in calculating the delay cost, where a large value signifies that the tenant’s workloads are more delay tolerant.

In addition, the server energy cost of tenant $k$ is given as follows:

$$E_{kj} = \kappa(s_{kj})^\gamma \lambda_k C_k,$$  

(10)

where $\kappa$ is the effective switched capacitance that depends on the chip architecture [20]. In this paper, the static energy due to motherboard, storage, memory, and network interface is separated from the model, because it is incurred regardless of the CPU decision.

IV. PROBLEM FORMULATION

We assume that each tenant can control and determine the CPU cycle at the same time as the announcement of the winning bids. As illustrated in Fig. 2, in this study we adopt a hierarchical optimization structure for bandwidth purchasing and server provision. At the beginning of each time slot $t$, a set of tenants requests bandwidth from the network provider. The tenant submit bids to the network provider that specify the bandwidth acquisition amount and the corresponding valuation. The true valuation is determined based on the delay cost and energy cost corresponding to the bandwidth amount. Based on the proposed auction-based scheme, the network provider can determine the winning bids and the corresponding payment. Tenants provide payment to the network provider and are allocated the amount of bandwidth declared in the winning bid. Then, based on the allocated bandwidth amount, tenants control the CPU cycle of their servers to minimize the cost due to delays and energy usage, which are related to the bidding valuation. We assume that each time slot is sufficiently long compared to the task execution time. Because the same process is repeated in each time-slot, we focus on one time-slot and ignore the time index. In section IV-A, we present the deciding tenants’ bids and the auction of the network provider.
A. Deciding tenants’ bids

We assume that each tenant can control and determine the CPU cycle at the same time as the announcement of the winning bids. Given a bandwidth amount, each tenant can determine the server provisioning to minimize the cost due to delays and energy consumption. Thus, the following represents the problem of tenant \( k' \) deciding the server provisioning with the given bandwidth so that it can minimize the total cost:

\[
\textbf{P1:} \quad \min_{s_{kj}} C_{kj} + \rho_k E_{kj} \\
\text{s.t.} \quad s_{kj} \in [0, s_{\text{max}}], \quad \forall k \in \mathcal{K},
\]

where \( \rho_k \) is the weight of the delay cost and energy cost of tenant \( k \).

In other words, for each bandwidth amount declared in a bid, the tenant can determine how much disutility it receives. The disutility (or total weighted cost) measures the tenant’s perception of its cost as a function of its bandwidth and CPU speed. The more disutility the tenant receives, the less the valuation is. Thus, we use negative cost as the valuation of a tenant. For example, for a given bandwidth amount declared in a bid, the valuation of the bid can be the inverse of the disutility (i.e., the weighted cost in (11)).

B. Auction

The goal of the auction is to maximize social welfare, which is defined as follows:

**Definition 1 (Social Welfare):** Social welfare in the bandwidth market is the aggregation of the tenants’ utility \( \sum_{k=1}^{K} \sum_{j=1}^{J_k} V_{kj} x_{kj} - \sum_{k=1}^{K} \sum_{j=1}^{J_k} p_{kj} x_{kj} \) and the network provider’s utility \( \sum_{k=1}^{K} \sum_{j=1}^{J_k} p_{kj} x_{kj} \). Payments between tenants and the network provider cancel out, and the social welfare is equal to \( \sum_{k=1}^{K} \sum_{j=1}^{J_k} V_{kj} x_{kj} \). Under the assumption of truthful bidding, the winner determination problem can be formulated as follows:

\[
\textbf{P2:} \quad \max_{x} \sum_{k=1}^{K} \sum_{j=1}^{J_k} b_{kj} x_{kj} \\
\text{s.t.} \quad (2), (3), (4).
\]

**Remark.** In our paper, for a given bandwidth amount declared in a bid, the valuation of the bid can be the inverse of the disutility (i.e., the weighted cost in (11)). By evaluating the cost corresponding to the renting bandwidth and server CPU cycle, the tenant comprehensively estimate the impact of renting bandwidth, taking into account not only the energy consumption but also the delay requirement. This enables the tenant to balance the energy consumption and the QoS of the server in terms of delay. In addition, the social welfare of this colocation MEC system is the sum of the valuations of bids submitted by tenants. By considering the valuation term into the tenant utility, we can see that the tenants try not only to obtain the high payment but also help the colocation MEC system achieve higher social welfare. Bandwidth allocation with higher social welfare indicates that the tenants provide more valuable services (lower energy consumption and/or lower delay) for application.

Maximizing social welfare is a standard and widely-considered goal in mechanism design, ensuring that resources are allocated in a Pareto optimal manner. In practice, social welfare maximization is a suitable goal for government-run edge data centers or for network providers that wish to optimize theirs bandwidth allocation for the best service quality rather than the highest revenue in a competitive market. However, this winner determination problem (WDP) is NP-hard.

**Theorem 1.** \( \textbf{P2} \) is NP-hard.

**Proof.** First, we prove that \( \textbf{P2} \) is an NP-problem. We assume that we have a solution instance, the tenant set \( \mathcal{K} \), all the bids submitted to the system, \( T_k, \forall k \in \mathcal{K} \), the total bandwidth \( W \), and the value of social welfare \( \Pi \). It takes a running time of \( O(K J_k) \) to verify this instance, which involves verifying whether

1) \( \sum_{k=1}^{K} \sum_{j=1}^{J_k} b_{kj} x_{kj} = \Pi \); 2) \( \sum_{j=1}^{J_k} x_{kj} \leq 1, \forall k \in \mathcal{K} \); 3) \( \sum_{k=1}^{K} \sum_{j=1}^{J_k} p_{kj} x_{kj} \leq W \). Because this verification requires polynomial time, \( \textbf{P2} \) is an NP problem.

To complete the proof that \( \textbf{P2} \) is NP-hard, we demonstrate that problem \( \textbf{P2} \) can be mapped to a known NP-hard problem in polynomial time. We consider a special case of \( \textbf{P2} \) in which each tenant submits one bid. This results in an instance \( A_1 = (J_k = 1, b_{kj} = l_k, \sum_{j=1}^{J_k} n_{kj} = \omega_k) \), which can map to an instance \( A = (l_1, \ldots, l_K, \omega_1, \ldots, \omega_K, K, W) \) of the knapsack problem given as

\[
\begin{align*}
\max \quad & \sum_{k=1}^{K} l_k x_k \\
\text{s.t.} \quad & \sum_{k=1}^{K} \omega_k x_k \leq W, \\
& x_k \in \{0, 1\}.
\end{align*}
\]

The mapping between \( A \) and \( A_1 \) takes polynomial time. Because the knapsack problem is a well-known NP-hard problem, problem \( \textbf{P2} \) is also NP-hard.

Because \( \textbf{P2} \) is NP-hard, the direct application of the VCG mechanism is computationally infeasible. This motivates us to design a randomized auction-based mechanism that can guarantee approximate truthfulness and be computationally efficient.

V. RANDOMIZED AUCTION BASED BANDWIDTH PURCHASING

In this section, we present the construction of the randomized auction mechanism [17], [18], [41]–[43]. We first present an overview of the randomized mechanism; then, we describe the fractional VCG mechanism and its solution. Next, we define an integrality gap and discuss the way to obtain it. Finally, we propose a scheme for allocating the bandwidth and a scheme for the payment determination of each winning bid.

A. Overview

There are four steps in constructing the randomized auction based mechanism. To design a truthful mechanism that can solve \( \textbf{P2} \) in polynomial time, we first move \( \textbf{P2} \) into the
Algorithm 1: Randomized Auction Algorithm

1. Step 1: Fractional VCG Auction
2. Step 2: Obtain the integrality gap \( \alpha \)
3. Step 3: Winner Selection
4. Step 4: Payment Determination

Let \( x \) denote the optimal fractional solution of \( \text{VCG} \) auction. The output is the optimal fractional solution \( x^* \) of each bid. Second, we calculate the integrality gap \( \alpha \) through an approximation algorithm. Third, we perform winner selection by decomposing the fractional VCG solution \( x^* \) into integer solutions \( x^\beta \) with probability weights \( \beta^\phi \). In this way, at the beginning of each time slot, an integer solution \( x^\beta \) is selected with probability \( \beta^\phi \). Before discussing the four steps of the randomized auction in detail, we provide the definition of fractional domain by relaxing (2) in \( \text{P2} \) to \( x_{kj} \geq 0, \forall k \in K, j \in J_k \). It should be noted that \( x_{kj} \leq 1, \forall k \in K, j \in J_k \) is ignored because of (3). \( \text{LPR} \) denotes the relaxed linear problem, and the VCG mechanism is then applied. This first step is called the fractional VCG mechanism. However, the optimal winner and payment solution of the fractional VCG mechanism are infeasible.

The following step helps to decompose the optimal fractional solution. This step requires knowledge of the integrality gap \( \alpha \), which is defined as follows.

Definition 1. Let \( \mathcal{P} \) denote the feasible solution space of \( \text{LPR} \) and \( \mathcal{Z} \subset \mathcal{P} \) denote the integer solution set of \( \text{P2} \). The integrality gap \( \alpha \) is as follows:

\[
\alpha = \sup \left( \frac{\max_{x \in \mathcal{P}} \sum_{k=1}^{K} \sum_{j=1}^{J_k} b_{kj} x_{kj}}{\max_{x \in \mathcal{Z}} \sum_{k=1}^{K} \sum_{j=1}^{J_k} b_{kj} x_{kj}} \right).
\]

Given the winner and payment determination of the fractional VCG mechanism and the value of \( \alpha \), we can perform winner selection and payment determination as follows: \( \alpha \) is used to decompose the fractional VCG allocation into a feasible integer solution of \( \text{P2} \) and scale down the fractional VCG payment. The four steps of the randomized auction are presented in Algorithm 1. First, we compute the fractional VCG auction. The output is the optimal fractional solution \( x^* \) and the payment \( p_k^t \) of each bid. Second, we calculate the integrality gap \( \alpha \) through an approximation algorithm. Third, we perform winner selection by decomposing the fractional VCG solution \( x^* \) into integer solutions \( x^\beta \) with probability weights \( \beta^\phi \). In this way, at the beginning of each time slot, an integer solution \( x^\beta \) is selected with probability \( \beta^\phi \).

Algorithm 2: The Greedy Approximation Algorithm

The three desired properties of randomized auction design as follows:

- Individual rationality: Each bidder obtains a non-negative utility by participating in the auction.
- Truthfulness: An auction is truthful in expectation if for any bid \( t_{kj} = \{n_{kj}, b_{kj}\} \), disclosing the true valuation is the dominant strategy, or the following requirement is met: \( E(\mu_{kj}(V_{kj}, t_{kj})) \geq E(\mu_{kj}(b_{kj}, -t_{kj})) \), where \( E(\mu_{kj}(V_{kj}, -t_{kj})) \) is the expected utility of \( t_{kj} = \{n_{kj}, b_{kj}\} \) and \( -t_{kj} \) as the set of bids excluding bid \( t_{kj} \).
- Approximate truthfulness: An auction is approximately truthful in expectation if for any bid \( t_{kj} = \{n_{kj}, b_{kj}\} \), the following condition is met: \( L(E(\mu_{kj}(V_{kj}, -t_{kj}))) \geq L(E(\mu_{kj}(b_{kj}, -t_{kj}))) \), where \( L(E(\mu_{kj}(V_{kj}, -t_{kj}))) \) is the lower bound of the expected utility of \( t_{kj} = \{n_{kj}, b_{kj}\} \). The concept of approximate truthfulness in this paper is the same as that in [43]. Approximate truthfulness is believed to be sufficient to guarantee truthfulness [43].
- Computational efficiency: Algorithms for winner selection and payment calculation in the auction run in polynomial time.

B. Computing the fractional VCG auction

Because the social welfare maximization problem \( \text{P2} \) is NP-hard, the optimal integer solution to this problem cannot be computed in polynomial time. By using the fractional VCG mechanism, the new problem \( \text{LPR} \) is linear programmable and can be solved in polynomial time. The fractional VCG mechanism consists of two steps. In the first step, we compute...
the optimal fractional solution \( x^* \) of LPR as follows:

\[
LPR : \max_{x \geq 0} \sum_{k=1}^K \sum_{j=1}^{J_k} b_{kj} x_{kj} \tag{15}
\]

\[
s.t. \quad (3), (4).
\]

In the second step, we compute fractional VCG payments for each bid according to the fractional VCG payment scheme:

\[
p^f_{kj} = M_T(t_{kj}) - \sum_{t_{ht} \in T(t_{kj})} b_{ht} x^*_{ht}, \tag{16}
\]

where \( T = \{T_k, k \in K\} \) is the set of all bids submitted to the network provider, and \( M_T(t_{kj}) \) is the maximum social welfare of \( P2 \) when bid \( t_{kj} \) is deleted from set \( T \). The sum \( \sum_{t_{ht} \in T(t_{kj})} b_{ht} x^*_{ht} \) is the social welfare except for the utility of bid \( t_{kj} \). Although the VCG mechanism can guarantee truthfulness, the result of the fractional VCG mechanism is not practically applicable. Therefore, we require a decomposition tool to obtain binary allocation. We thus propose a greedy approximation algorithm to achieve the integrality gap, which is then used in allocation decomposition.

**C. Greedy approximation algorithm**

In this subsection, we derive the upper bound of \( \alpha \) using the dual fitting method. We formulate the dual of LPR by introducing dual variables \( \phi \) and \( \psi \) for constraints (3) and (4), respectively, as follows:

\[
\min_{\phi, \psi} \sum_{k=1}^K \phi_k + W \psi \tag{17a}
\]

\[
s.t. \quad \phi_k + n_{kj} \psi \geq b_{kj}, \quad \forall k, j, \tag{17b}
\]

\[
\phi_k, \psi \geq 0, \quad \forall k. \tag{17c}
\]

In the following, we introduce Algorithm 2, which computes the approximation solution of \( P2 \), and we then determine \( \alpha \) via dual fitting by theoretical analysis. Algorithm 2 iteratively constructs the feasible integer solution to \( P2 \) and solution to the dual of LPR by selecting bids with the highest unit-weight until all tenants are satisfied or the network provider does not have sufficient bandwidth to allocate to all tenants.

**Theorem 2.** Algorithm 2 provides a feasible solution to \( P2 \) and LPR.

The proof is provided in Appendix A.

**Lemma 1.** In Algorithm 2, \( \phi \) and \( \psi \) are feasible solutions for the dual.

The proof is provided in Appendix B.

**Theorem 3.** The upper bound of integrality gap \( \alpha \) between \( P2 \) and LPR and the approximation ratio of our proposed greedy approximation algorithm is \( 1 + \frac{W}{W - R} \), where \( R \) is the maximum bandwidth demand among submitted bids and \( \epsilon \) is the maximum ratio of bandwidth demands among bids submitted by the same tenant.

The proof is provided in Appendix C.

**D. Winner Selection**

Using the upper bound of \( \alpha \) obtained in the previous step, we perform winner selection by decomposing the fractional VCG auction result into a set of feasible integer solutions. Specially, we identify the non-negative multipliers \( \beta = \{\beta^q\} \), such that \( \sum_{q \in Q} \beta^q x^q \geq x^* / \alpha \) and \( \sum_{q \in Q} \beta^q = 1 \), where \( Q \) is the index set for all feasible integer solutions to \( P2 \). Based on the decomposition tool [44], values for \( \beta^q \) are computed by solving the following problem:

**Primal:**

\[
\min_{\beta} \sum_{q \in Q} \beta^q \tag{18}
\]

\[
s.t. \quad \sum_{q \in Q} \beta^q x^q = x^*, \quad \sum_{q \in Q} \beta^q \geq 1, \quad \beta^q \geq 0, \quad \forall q \in Q.
\]

The first constraint of (18) describes the decomposition; thus, if the optimum satisfies \( \sum_{q \in Q} \beta^q = 1 \), we are almost done. However, (18) has an exponential number of variables; therefore, it takes time to solve (18) directly. We transform (18) into the dual linear program as follows:

**Dual:**

\[
\max_{\omega, i} \sum_{k,j} \omega_{kj} x^*_{kj} / \alpha + i \tag{19}
\]

\[
s.t. \quad \omega_{kj} x^*_{kj} + i \leq 1, \quad \forall q \in Q,
\]

\[
i \geq 0, \quad \forall k.
\]

In (19), the number of constraints is exponential. If (19) is solved by the ellipsoid method, it can be solved in polynomial time. Specifically, \( \omega = \{\omega_{kj}\} \) is considered a valuation vector for bids, and a feasible integer solution is the output of Algorithm 2. Each feasible integer solution corresponds to each hyperplane that cuts the ellipsoid until the final optimal solution can be obtained. In addition, each feasible integer solution corresponds to variable \( \beta^q \) in the primal problem. The number of hyperplanes is polynomial; therefore, the primal problem has a polynomial number of variables. This allows the primal problem to be solved in polynomial time.

Next, we prove that the primal has the optimal value of \( \sum_{q \in Q} \beta^q = 1 \).

**Theorem 4.** The primal (18) and dual (19) can be optimally solved with an optimal objective value of 1.

**Proof.** See Lemma 3.5 in [18].

**E. Payment Determination**

We select \( x^q \) with probability \( \beta^q \), and the payment of each winning bid is decided as follows:

\[
p_{kj} = \begin{cases} \frac{x^q_{kj}}{x^*_{kj}}, & \text{if } x^*_{kj} \neq 0, x^q_{kj} \neq 0, \\ 0, & \text{otherwise.} \end{cases}
\]
VI. PROPERTIES OF THE PROPOSED RANDOMIZED AUCTION MECHANISM

Proposition 1. The proposed mechanism is approximately truthful in expectation.

Proof. The expected utility of bid $t_{kj}$ is as follows:

$$\begin{align*}
&b_{kj}(\sum_{q \in \mathcal{Q}} \beta^q x_{kj}^q) - E(p_{kj}) \\
&= b_{kj}x_{kj}^*\Theta_{kj} - (\sum_{q \in \mathcal{Q}} \beta^q x_{kj}^q p_{kj}^f/x_{kj}^*) \\
&= b_{kj}(x_{kj}^* \Theta_{kj}) - p_{kj}^f \Theta_{kj} \\
&= (b_{kj}(x_{kj}^* - p_{kj}^f))\Theta_{kj} \\
&= (b_{kj}(x_{kj}^* - p_{kj}^f))/\alpha
\end{align*}$$

(20)

where $\Theta_{kj} = (\sum_q \beta^q x_{kj}^q)/x_{kj}^* = 1/\alpha$. \hfill $\blacksquare$

It can be seen that the lower bound of the expected utility for each bid of each tenant is $1/\alpha$ times the lower bound of the utility in the fractional VCG mechanism. The fractional VCG mechanism is truthful; thus, the randomized auction is approximately truthful.

Proposition 2. The proposed mechanism is individually rational in expectation.

Proof. First, we verify the individual rationality of the fractional VCG mechanism. Because $x_{kj}^*b_{kj} + \sum_{b_kr, t \in T|b_kr} x_{kjr}^*b_{kjr}$ is the maximum social welfare of (15), the following inequality is true:

$$\begin{align*}
x_{kj}^*b_{kj} + \sum_{t_kr, t \in T|t_kr} x_{kjr}^*b_{kjr} &\geq V_T|t_{kj}
\end{align*}$$

(21)

Therefore, we have

$$\begin{align*}
x_{kj}^*b_{kj} - (V_T|t_{kj} - \sum_{t_kr, t \in T|t_kr} x_{kjr}^*b_{kjr}) &\geq 0
\end{align*}$$

(22)

The left-hand side of (22) is the utility of bid $t_{kj}$ in the fractional VCG mechanism. Thus, the fractional VCG mechanism is individually rational. According to Proposition 1, the ratio of the expected utility of bid $t_{kj}$ in the randomized auction and the fractional VCG mechanism is $\Theta_{kj} = 1/\alpha$. Therefore, each bid in the proposed mechanism is individually rational in expectation. \hfill $\blacksquare$

Proposition 3. The randomized auction is computationally efficient.

Proof. Steps 1 and 4 are computationally efficient, because (15) is linearly programmable and takes polynomial time to solve. Step 2 also has a polynomial time complexity. In Algorithm 2, the termination conditions ensure that the while loop in Algorithm 2 iterates at most $K$ times, linear to the input size. Within the loop, lines 6-13 can be completed in polynomial time. Therefore, Algorithm 2 runs in polynomial time. In Step 3, the ellipsoid method is used to decompose the fractional VCG result into a set of feasible integer solutions. It then takes polynomial-time to complete Step 3 [45]. \hfill $\blacksquare$

VII. SIMULATION RESULTS

In this section, we discuss simulations performed to evaluate the proposed randomized auction-based bandwidth trading scheme. The CPU request is uniformly distributed within [0.1, 1] GHz [21]. The file size is uniformly distributed within [100, 500] kB. Other parameters are $D_{th} = 2$ ms, $\kappa = 10^{-28}$, $\gamma = 2$, $\rho_k = 1$. The total bandwidth is 10 MHz. The coverage area (cell radius) of the BS is 1000 m, and the transmission power is 20 W. We consider a log-distance path loss model provided by [46]. The power density of the noise is $-174$ dBm/Hz.

A. Performance of the Approximation Algorithm

To demonstrate the functionality of the approximation algorithm, we first evaluate the performance of Algorithm 2 in terms of average social welfare, approximation ratio, and running time by varying the number of tenants, the demand bandwidth (bandwidth bidding amount) and the available dynamic bandwidth (available bandwidth after static allocation). We compare the proposed greedy algorithm with three baselines:

- OPT: the optimal solution is obtained by solving P2 optimally, which can obtained by matlab solver.
- Alg.3 [18]: the solution is obtained based on [18, Algorithm 3]
- Random: the winners are selected randomly so that the total bandwidth demand of the winners does not exceed the available bandwidth.

We sequentially consider the three following scenarios. In Scenario 1, the number of tenants varies from 60 to 100 with a step size of 10. Each tenant can submit five bids. The demand bandwidth is uniformly distributed between [2%, 4%] of the total bandwidth, and the available bandwidth is randomly generated between [50%, 70%] of the total bandwidth. In Fig. 3a, the result of the proposed greedy approximation algorithm is closer to the optimal solution compared with the random scheme. When the number of bids is 500, the average social welfare by the proposed greedy approximation algorithm is 93.75% of the optimal solution. When bid number is 500, the proposed greedy approximation algorithm can improve the average social welfare obtained by Alg.3 [18] by 3.2%. Furthermore, social welfare increases as the number of bids increases, which can be observed in OPT, the proposed greedy approximation algorithm and Alg.3 [18]. The reason for this is that when the number of bids increases, the network provider has larger number of chances to select the bid with a higher valuation so that the network provider obtains higher social welfare. In Scenario 2, the number of tenants is fixed at 100, and each tenant submits five bids. The available dynamic bandwidth is randomly generated between [50%, 70%] of the total bandwidth while the demand bandwidth is uniformly distributed between [0%, u1%] of the total bandwidth, where $u_1$ is increased from 2% to 4% with a step size of 0.5%. Fig. 3b illustrates the increase in the average social welfare when the demand dynamic bandwidth increases in three schemes: OPT, Alg. 3 [18] and the proposed algorithm when the demand dynamic bandwidth increases from 2% to 2.5%. In Appendix
welfare when the available dynamic bandwidth increases. This is due to the fact that when there are more available dynamic resources, more bidders are selected as winners, resulting in higher average social welfare. In all three scenarios, the random scheme is the poorest because the winning bids are selected without considering the valuation of the bids.

In the following, we study the complexity performance over the number of bids submitted by all tenants. We compare the running time of the proposed greedy approximation algorithm with the optimal solution to solve the original NP-hard problem and Alg. 3 [18]. As seen in Table II, our proposed solution is much simpler than the optimal solution, especially when the number of bids increases. However, it takes more running time than Alg. 3 [18] but this running time difference is not significant.

We also evaluate the theoretical and real approximation ratios. Fig. 4a demonstrates that the real approximation ratio and the theoretical ratio proved in Theorem 3. We observe that when the number of tenants are 10, 20, 30, 40, 50, the median of the real approximation ratio are 1.0007 and 1.1251 while the lowest and the highest valuation of real approximation ratio are 1.0575, 1.0611, respectively and the median of the theoretical ratio are 3.0155, 3.0529, 3.0749, 3.0806, 3.0914, respectively. Moreover, we can also see the lowest and the highest valuation of real approximation ratio are 1.0007 and 1.1251 while the lowest and the highest valuation of real approximation ratio are 2.782 and 3.1626. Therefore, we conclude that the real approximation ratio is much lower than the theoretical ratio.

B. The Randomized Auction

We implement the randomized auction by applying the ellipsoid method and the approximation algorithm. Fig. 5a compares the social welfare achieved by the fractional VCG mechanism, our proposed randomized auction, the randomized auction [18], and the lower bound of the expected social welfare. Although the social welfare achieved by the fractional VCG auction is the highest among four schemes, the fractional VCG is actually infeasible in practice. The line corresponding to lower bound of the expected social welfare in Fig. 5a is equal to 1/α times the social welfare obtained in the fractional
Fig. 5: Numerical results for a) social welfare of the randomized auction compared with the fractional VCG auction, and b) total payment of the randomized auction compared with the fractional VCG auction.

Fig. 6: Utility of the first bid of Tenant 1 under different bid ratios.

Fig. 7: Expected achieved valuation and payment of bids.

VCG mechanism. Our proposed randomized auction provides social welfare that is higher than the randomized auction in [18] and higher than the lower bound of the expected social welfare line. As the tenant number is 100, the social welfare achieved by our proposed randomized auction is 2% higher than the social welfare achieved by the randomized auction in [18]. Fig. 5b presents the corresponding total payments.

Fig. 8: Average bandwidth utilization achieved by 2 mechanisms: proposed randomized auction and the randomized auction [18].

Fig. 6 illustrates the approximate truthfulness performance. The valuation ratio on the x-axis is defined as the ratio of the valuation of the submitted bid and the true valuation. In Fig. 6, the bidder can achieve maximum utility when the valuation ratio is 1. When the bidder cheats, the utility is lower and may even be negative. Fig. 7 illustrates the performance of the individual rationality of the proposed randomized auction. We consider five tenants, each of which can submit three bids to the network provider. As seen in Fig. 7, the expected achieved valuation of each bid is higher than the payment; therefore, the utilities of the tenants are non-negative.

Fig. 8 presents a comparison between the proposed randomized auction and the randomized auction [18] in terms of the bandwidth utilization. The bandwidth utilization is calculated as the proportion of the allocated bandwidth. Fig. 8 demonstrates that the utilized bandwidth in the proposed randomized auction is 3.1% better than in the randomized auction [18] when the number of tenants is 90 (bid number is 180). The reason for this is that the stopping criteria in our proposed greedy approximation algorithm are more flexible than in Alg. 3 [18].
VIII. CONCLUSION

In this paper, we study bandwidth trading between a network provider and tenants in a colocation MEC system. We propose a trading bandwidth mechanism in which the network provider acts as an auctioneer and leases bandwidth to the tenants. The network provider can provide incentives to the tenants to participate in the bandwidth trading market with truthful information. Each tenant submits its bid containing valuation and bandwidth demand. In addition, we propose a randomized auction mechanism for the social welfare maximization problem, which is proven to be NP-hard. An approximation algorithm for winner selection is provided with polynomial-time complexity and is used as an element unit to realize the randomized auction mechanism. Simulation results reveal that the proposed randomized auction mechanism offers approximate truthfulness and individual rationality, and achieves better performance than the baseline in terms of average social welfare. In particular, the proposed mechanism achieves 2% higher average social welfare compared to the existing randomized auction when the number of tenants is 100. With the guidance supplied in [27], we plan to extend our work by integrating machine learning (e.g., deep Q-learning neural network) to predict the dynamic changes of user demand or unexpected server failure and make the appropriate modifications beforehand (e.g., borrow the extra resources from the other tenants before a peak in usage or server failure).

APPENDIX A
PROOF OF THEOREM 2

Proof. First we determine whether the returned solution is a feasible solution to P2. Values in x are initialized to 0 and updated to 1 only in iterations. Thus constraints (2) and (4) are satisfied. The proposed greedy approximation algorithm is terminated when either one of the following is satisfied: the network provider has insufficient bandwidth to support the demand of tenants (line 13), or all tenants have won one bid (line 5). Therefore, constraint (3) is not violated. We can verify that a feasible solution to P2 is also a feasible to its LPR.

APPENDIX B
PROOF OF THE LEMMA 1

Proof. We discuss the following three cases:

- Case 1: tenant \( µ \) wins, that is \( µ ∈ \mathcal{U} \) and \( b_{µj} = \max_{j' ∈ \mathcal{J}_µ} \{ b_{j'j} \} \). Then we have \( φ_µ = b_{µj} ≥ b_{j'j}, ∀ j' ∈ \mathcal{J}_µ \). Thus, constraint (17b) is satisfied for all tenants in \( \mathcal{U} \).

- Case 2: tenant \( µ \) loses the auction, that is \( µ ∈ \mathcal{K} \setminus \mathcal{U} \). According to the while loop, it is evident that

\[
\frac{b_{kj}}{n_{kj}} > \frac{b_{j'j}}{n_{j'j}}, ∀ k ∈ \mathcal{U}.
\]

Therefore, \( \psi > \frac{b_{j'j}}{n_{j'j}} \). Thus,

\[
\tilde{\psi} ≥ \frac{b_{j'j}}{n_{j'j}} ≥ \frac{b_{j'j}}{n_{j'j}}.
\]

In addition, we have

\[
b_{j'j} ≥ b_{j'j} \quad \text{and} \quad \epsilon > \frac{n_{j'j}}{n_{j'j}}, ∀ j' \neq j_µ.
\]

Therefore,

\[
\tilde{\psi} ≥ \frac{b_{j'j}}{n_{j'j}}, ∀ j' \neq j_µ.
\]

Therefore, constraint (17b) is also satisfied for all tenants in \( \mathcal{K} \setminus \mathcal{U} \).

APPENDIX C
PROOF OF THEOREM 3

Proof. In Lemma 1, we prove that \( \phi \) and \( \tilde{\psi} \) are feasible solutions for the dual formulation. Therefore, according to LP duality, \( \sum_{k=1}^{K} ϕ_k + W \psi \) is the upper bound for the primal formulation LPR. Let OPT and OPT\( _f \) be the optimal solution for (12), (15). We can obtain the following:

\[
\text{OPT} ≤ \text{OPT}_f ≤ \sum_{k=1}^{K} ϕ_k + W\tilde{\psi}
\]

\[
≤ \sum_{k∈\mathcal{U}} b_{kj} + W\tilde{\psi}
\]

\[
≤ \sum_{k∈\mathcal{U}} b_{kj} + W\epsilon + \sum_{k∈\mathcal{U}} b_{kj}
\]

\[
≤ \sum_{k∈\mathcal{U}} b_{kj} (1 + \frac{W\epsilon}{W − R})
\]

\[
≤ p(1 + \frac{W\epsilon}{W − R})
\]

where \( R = \max_{k,j} n_{kj} \) and \( p \) is the social welfare of the near-optimal integer solution computed by the proposed greedy approximation algorithm. Therefore, the integrality \( α \) is given as

\[
\frac{\text{OPT}_f}{\text{OPT}} ≤ \frac{\text{OPT}_f}{p} \leq (1 + \frac{W\epsilon}{W − R})
\]

The approximation ratio is

\[
\frac{\text{OPT}}{p} ≤ \frac{\text{OPT}_f}{p} ≤ (1 + \frac{W\epsilon}{W − R})
\]

APPENDIX D
MONOTONICITY PROPERTY OF BID VALUATION

Lemma 2. If two bids of tenant \( k \) are \( n_{kj} \) and \( n_{kh} \) such that \( n_{kj} > n_{kh} \), we have the corresponding payment \( V_{kj} > V_{kh} \).
Proof. We have $R_{kj} > R_{th}$ due to $n_{kj} > n_{th}$. Hence, $D^t_{kj} < D^t_{th}$. Following the equation (9) and (10), we have

$$C_{kj} + \varphi_k E_{kj} = \begin{cases} \beta_k(D_{kj} - D_{th}) + \varphi_k \kappa (s_{kj})^\gamma \lambda_k, & \text{if } D_{kj} > D_{th}, \\ \varphi_k \kappa (s_{kj})^\gamma \lambda_k, & \text{otherwise.} \end{cases}$$

Set $Q_{kj}^1 = \beta_k(D_{kj} - D_{th}) + \varphi_k \kappa (s_{kj})^\gamma \lambda_k$ and $Q_{kj}^2 = \varphi_k \kappa (s_{kj})^\gamma \lambda_k$. After getting the first order derivative of $Q_{kj}^1$, this function gets the minimum value at the point where $s^*_{kj} = \frac{1}{1+\sqrt{\frac{\beta_k}{\varphi_k \kappa}}}$ The condition $D_{kj} > D_{th}$ is equivalent to $s_{kj} < \frac{1}{D_{th} - D_{th}^t}$. In addition, $Q_{kj}^2$ is the increasing function of $s_{kj}$. We also have $Q_{kj}^1\left(\frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}\right) = Q_{kj}^2\left(\frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}\right)$. Similarly, we set $Q_{kj}^1 = \beta_k(D_{kj} - D_{th}) + \varphi_k \kappa (s_{kj})^\gamma \lambda_k$ and $Q_{kj}^2 = \varphi_k \kappa (s_{kj})^\gamma \lambda_k$. The optimal point $Q_{kj}^2$ is $s^*_{kj} = \frac{1}{1+\sqrt{\frac{\beta_k}{\varphi_k \kappa}}}$ in the condition $s_{kj} < \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}$.

There are three cases to consider:

- **Case 1**: $s_{max} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}$, follow the above.

  **Case 1.A:** $s_{max} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}$, follow the above.

  Thus, $V_{kj} = Q_{kj}^1(s_{kj}) = \frac{1}{Q_{kj}^2(s_{kj})}$.

- **Case 2**: $\beta_k C_{kj} < D_{th} - D_{th}^t$, follow the above.

  **Case 2.A:** $\frac{\beta_k C_{kj}}{D_{th} - D_{th}^t} > s_{max} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}$, this subcase is similar to the subcase 1.A. Therefore, we have $V_{kj} > V_{th}$.

  **Case 2.B:** $\frac{\beta_k C_{kj}}{D_{th} - D_{th}^t} > s_{max} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}$, this subcase is similar to the subcase 1.C. Therefore, we have $V_{kj} > V_{th}$.

  **Case 2.C:** $\frac{\beta_k C_{kj}}{D_{th} - D_{th}^t} > s_{max} > s_{max} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}$. Therefore, $V_{kj} = \frac{1}{Q_{kj}^2(s_{max})}$ and $V_{th} = \frac{1}{Q_{kj}^1(s_{max})}$.

- **Sub Case 2.D**: $s_{kj}^* = s_{kj}^* > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t} > s_{max} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}$, Therefore, $V_{kj} = Q_{kj}^1\left(\frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}\right)$ and $V_{th} = \frac{1}{Q_{kj}^2(s_{max})}$. However, we have $Q_{kj}^2\left(\frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}\right) > Q_{kj}^1\left(\frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}\right)$. Thus, $V_{kj} > V_{th}$.

- **Case 3**: $\frac{\beta_k C_{kj}}{D_{th} - D_{th}^t} > s_{max} > \frac{\beta_k C_{kj}}{D_{th} - D_{th}^t}$. This subcase is similar to the subcase 1.A. Therefore, we have $V_{kj} > V_{th}$.


**REFERENCES**


Thant Zin Oo received B.S. degree in computing and information system from London Metropolitan University, U.K., in 2008, the B.Eng. degree in electrical systems and electronics from Myanmar Maritime University, Thanlyin, Myanmar, in 2008, and the Ph.D. degree in computer science and engineering from Kyung Hee University, South Korea, in 2017. He received scholarships from Kyung Hee University for his graduate studies and from British Council for his undergraduate studies. His research interests include wireless communications, network virtualization, data centers, sustainable energy, optimization techniques, and artificial intelligence.

Ki Tae Kim received the B.S. and M.S. degrees in computer science and engineering from Kyung Hee University, Seoul, South Korea, in 2017 and 2018, respectively, where he is currently pursuing the Ph.D. degree in computer science and engineering. His research interests include SDN/NFV, wireless networks, unmanned aerial vehicle communications, and machine learning.

Choong Seon Hong (S’95-M’97-SM’11) received the B.S. and M.S. degrees in electronic engineering from Kyung Hee University, Seoul, South Korea, in 1983 and 1985, respectively, and the Ph.D. degree from Keio University, Tokyo, Japan, in 1997. In 1988, he joined KT, Gyeonggi-do, South Korea, where he was involved in broadband networks as a member of the Technical Staff. Since 1993, he has been with Keio University. He was with the Telecommunications Network Laboratory, KT, as a Senior Member of Technical Staff and as the Director of the Networking Research Team until 1999. Since 1999, he has been a Professor with the Department of Computer Science and Engineering, Kyung Hee University. His research interests include future Internet, intelligent edge computing, network management, and network security.

Dr. Hong is a member of the Association for Computing Machinery (ACM), the Institute of Electronics, Information and Communication Engineers (IEICE), the Information Processing Society of Japan (IPSJ), the Korean Institute of Information Scientists and Engineers (KIIEEE), the Korean Institute of Communications and Information Sciences (KICS), the Korean Information Processing Society (KIPS), and the Open Standards and ICT Association (OSIA). He has served as the General Chair, the TPC Chair/Member, or an Organizing Committee Member of international conferences, such as the Network Operations and Management Symposium (NOMS), International Symposium on Integrated Network Management (IM), Asia-Pacific Network Operations and Management Symposium (APNOMS), End-to-End Monitoring Techniques and Services (E2EMON), IEEE Consumer Communications and Networking Conference (CCNC), Assurance in Distributed Systems and Networks (ADSN), International Conference on Parallel Processing (ICPP), Data Integration and Mining (DIM), World Conference on Information Security Applications (WISA), Broadband Convergence Network (BoC), Telecommunication Information Networking Architecture (TINA), International Symposium on Applications and the Internet (SAINT), and International Conference on Information Networking (ICOIN). He was an Associate Editor of the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT and the IEEE JOURNAL OF COMMUNICATIONS AND NETWORKS and an Associate Editor for the International Journal of Network Management and an Associate Technical Editor of the IEEE Communications Magazine. He currently serves as an Associate Editor for the International Journal of Network Management and Future Internet Journal.

Shaolei Ren is an Associate Professor of Electrical and Computer Engineering at the University of California, Riverside, where he also holds a Cooperating Faculty position in the Department of Computer Science and Engineering. He received his B.E. from Tsinghua University in 2006, M.Phil. from the Hong Kong University of Science and Technology in 2008, and Ph.D. from the University of California, Los Angeles in 2012. His research interests broadly include networks, optimization and learning. He was a recipient of the U.S. NSF Faculty Early Career Development (CAREER) Award in 2015.