International Journal of Innovative
Computing, Information and Control
Volume **17**, Number **5**, October **2021**

ICIC International ©2021 ISSN 1349-4198
pp. **1581–1598**

OPTIMAL HETEROGENEOUS CLOUDLET PLACEMENT WITH DELAY CONSTRAINTS AND USER DYNAMICS IN WMANS

Hengzhou Ye*, Fengyi Huang and Wei Hao

Guangxi Key Laboratory of Embedded Technology and Intelligent System
Guilin University of Technology
No. 319, Yanshan Street, Yanshan District, Guilin 541006, P. R. China
*Corresponding author: 2002018@glut.edu.cn; { 2120190614; 1020190598 }@glut.edu.cn

Received April 2021; revised July 2021

Abstract. *Cloudlet is proposed to place the data center at the edge of the network to ensure the quality of service (QoS) for delay-sensitive tasks in mobile edge computing (MEC). Mobile users can randomly roam in the network area, which lead to different density and resource demands for users in wireless metropolitan area network (WMAN). Considering users' QoS, the appropriate wireless access points (APs) are selected in WMAN for cloudlet deployment, which can greatly reduce user access delay. From the perspective of infrastructure service providers (ISP), it is critical to place heterogeneous cloudlets according to different regional user resource demands in WMAN. In this article, both users' mobility and heterogeneous cloudlets with limited resource capacity are considered in WMAN, therefore, we firstly design a heterogeneous cloudlet deployment model aiming at minimizing the number of cloudlet deployments, which is formulated as an integer linear programming problem (ILP), and the improved heuristic algorithm for minimizing the number of cloudlet deployments (IMHA) is thus proposed to solve it. Multi-dimensional simulations are conducted to show that the improved model and IMHA can not only effectively reduce the average transmission delay, but also reduce the number of cloudlet deployments while meeting the users' tolerable access delay.*
Keywords: Mobile edge computing (MEC), Heterogeneous cloudlet deployment, Minimization, Quality of service (QoS), Average transmission delay

1. **Introduction.** Due to the limited computation storage and battery resource capabilities of mobile devices, mobile devices can offload task requests to remote cloud data centers to solve the resource limitations of portable mobile devices and extend battery life through mobile cloud computing (MCC) [1]. However, remote cloud data centers are usually far away from mobile users, which will increase the communication delay between mobile devices and remote clouds [2-4], and bring greater challenges to delay-sensitive and intensive mobile applications, such as augmented reality (AR), virtual reality (VR) and online games [5-8]. In order to solve the problem of higher latency between mobile users and remote cloud data centers, pioneer researchers proposed the concept of MEC, which is a key technology in the emerging fifth-generation (5G) network that can host computing-intensive applications, and the network in MEC is very close to mobile users by placing small cloud server infrastructure with limited resources such as cloudlet at the edge of the network, and provides context-aware services based on network information [9-12], cloudlet technology can effectively reduce the user's service access delay [13]. Therefore, the best location for cloudlet deployment is at the edge of the network closer to mobile devices, such as mobile base stations or APs [14].

DOI: 10.24507/ijicic.17.05.1581

Most existing researches focus on offloading user tasks to cloudlet [15-19]. However, the existing task offloading framework is an edge computing environment provided by the same ISP, which has poor scalability, and more mobile devices and different types of devices are not considered in the experiment to test the compatibility of the offloading framework. And when considering resource allocation problem in cloudlet environment, a load-aware resource allocation strategy is designed, which can adaptively allocate cloudlet resources to delay-sensitive mobile user task requests and solve the problem of users' QoS and budget of ISP [20-23]. However, for applications with strict accuracy requirements, it is not considered that the unreliable connection between the mobile device and the cloudlet may lead to dependency conflicts between applications or incorrect execution results. Yang et al. [24] developed an accurate mixed integer linear programming (MILP) formula and a benders decomposition algorithm to solve the task allocation problem which ignores the scenario of task queuing and task priority. In order to meet the required delay requirements, task migration [25-28] considers the problem of migrating tasks assigned to cloudlet to a nearby or remote cloud platform when cloudlet is overloaded. When considering cloudlet placement problem, the cloudlet placement framework in wireless networks with delay as the constraint is proposed [29-31]. However, existing strategies fail to consider the impact of the heterogeneity of cloudlet on deployment costs and the QoS of delay-sensitive tasks. Therefore, Zhao et al. [32] used software-defined networking (SDN) technology and proposed an enumeration-based optimal placement algorithm (EOPA) to minimize the average access latency and provide flexible and programmable management for cloudlet deployment. Guan et al. [33] considered the long-term cost of optimizing cloudlet deployment and operation. However, existing research does not fully consider the impact of cloudlet performance differences on the cost of cloudlet deployment. It mainly considers where cloudlets are deployed, while ignoring the issue of cloudlet computing resource levels.

By our literature survey, little attention has been paid to the minimization of the number of heterogeneous cloudlet deployments in WMAN. Although the cloudlet technology can reduce the end-to-end delay of users, the budget of the cloudlet provider is limited. Because users can roam in the network area and offload task requests to cloudlet, which lead to different user densities in different network areas, it is necessary to deploy a large number of cloudlets in dense areas of mobile user equipment for task offloading. If a small amount of cloudlets are deployed in dense areas, it will affect users' QoS. Therefore, it is not to be ignored to choose the cloudlet deployment location for cloudlet deployment and minimize the number of cloudlet deployments according to the density of users in different regions and the different user resource requirements. Similar to the problem of minimization of the number of network nodes, a grid-based clustering dynamic forwarding routing protocol (GCDF) is proposed to minimize the number of dead nodes in wireless sensor networks (WSN) [34]. Furthermore, K heuristic algorithm is proposed to solve the problem of cloudlet placement in a WMAN [35,36]. However, the minimization of K heuristic algorithm selects the optimal AP for the cloudlet deployment according to the order of the transmission delay between wireless APs, and ignores the mobility of users and the heterogeneity of cloudlet servers. Yao et al. [37] considered user mobility and first proposed the lowest cost of heterogeneous cloud deployment, while ensuring the QoS of users in the MEC environment, but it ignores the average latency of APs transmitting user requests and the resource demands of user tasks, and HA is proposed to solve this problem; however, this algorithm chooses the best AP according to APs degree, which cannot balance the workload of wireless APs and transmission cost between wireless AP points.

Therefore, the existing research on minimizing the number of heterogeneous cloudlet deployments in WMAN needs to be improved, because APs with heavy workloads may not be the closest to the user, which will increase the users' tolerable delay. Therefore, based on the research in [35,37], this paper considers the mobility of users, the delay of wireless APs transmitting user task requests, and the amount of resource demands for any user to access the AP, and design a model of the minimization of the number of heterogeneous cloudlet deployments, and IMHA based on the average delay of wireless APs transmitting user requests is proposed to minimize the number of heterogeneous cloudlet placements, which combine probability between users and wireless APs with the transmission delay the wireless AP transmitting the user request to cloudlet; therefore, the average transmission time is calculated to select the APs and choose the optimal AP for the cloudlet placement.

The main contributions of this article are as follows.

1) We have comprehensively considered the user dynamics, the heterogeneity of cloudlet and the delay of wireless AP transmitting user requests, and designed a new model of minimizing the number of cloudlet placements to study heterogeneous cloudlet deployment and user association issues; meanwhile, the problem is proved to be an NP-hard problem.

2) By improving the method of selecting the location of cloudlet deployment, IMHA is adopted to minimize the number of cloudlet deployments.

3) Finally, a large number of simulation experiments show that our improved model and algorithm are effective.

The rest of the paper is organized as follows. Preliminaries and problem statement are introduced in Section 2. Section 3 presents heterogeneous cloudlet placement algorithm design. Section 4 introduces experimental analysis. In the end, conclusions are shown in Section 5.

2. **Preliminaries and Problem Statement.**

2.1. **System model.**

As is shown in Figure 1, a WMAN system can consist of a set of wireless APs $V = \{v_1, v_2, \ldots, v_m\}$ connected to each other via the Internet, a set of potential locations S for cloudlet deployment, and a set of users $U = \{u_1, u_2, \ldots, u_n\}$ offloading task requests to cloudlet through wireless APs. An undirected graph $G = \{V \cup S \cup U, E\}$ is used to represent the relationship among mobile users, wireless APs and cloudlet placed with the wireless APs in the WMAN. E is used to represent the link set between two wireless APs or between the wireless AP and the cloudlet placed with the wireless AP. In a WMAN, users randomly roam in different network areas, and the location of each user may change over time, and wireless APs are deployed in places with relatively high population density, such as train stations, shopping malls, and bus stations. Therefore, in the WMAN, the number of requests received by each wireless AP v_k of any user u_i is closely related to the contact probability of the user with the wireless AP v_k, therefore, R_k is used to represent the set of task requests received by wireless AP v_k, and q_{ik} represents the contact probability of user u_i with the wireless AP v_k.

Therefore, the number of task requests from user $u_i \in U$ received by wireless AP $v_k \in V$ is determined by Equation (1).

$$Z_{ik} = R_k * q_{ik} \quad \forall i \in U \quad \forall k \in V \tag{1}$$

$Z(v_k)$ is used to represent the number of user task requests received by wireless AP v_k, which can be described as Equation (2).

$$Z(v_k) = \sum_{i \in U} R_k * q_{ik} \quad \forall k \in V \tag{2}$$

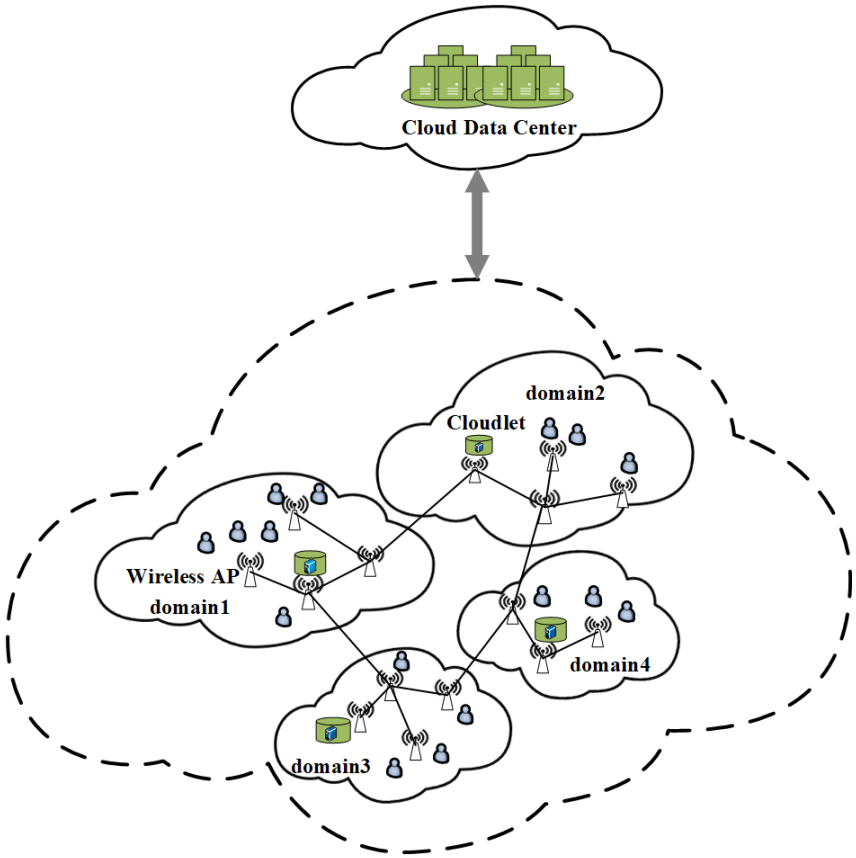

FIGURE 1. Cloudlet ($k = 4$) placement in the WMAN

Consequently, the number of user task requests received by all wireless APs V is calculated as Equation (3).

$$Z(V) = \sum_{k=1}^{|V|} Z(v_k) \quad \forall k \in V \tag{3}$$

It is assumed that the deployment location of the cloudlet is the same as the wireless AP, only one cloudlet can be deployed to each potential wireless AP, and k cloudlet servers need to be deployed to k different potential deployment locations in the set S. $F = \{f_1, f_2, \ldots, f_k\}$, $1 \leq k \leq |S|$ is denoted as a set of cloudlet server. For each cloudlet server f_i, $1 \leq i \leq k$, there are limited computing resources to process the allocated user requests, r_k is introduced to represent the resource capacity of f_k, we assume that cloudlet servers are heterogeneous, $f_i \neq f_j$, $r_i \neq r_j$. The processing capacity of the server is usually proportional to the resource capacity, and different cloudlet servers have different resource capacities [35]. If the AP area v_j where user u_i is located has been deployed with cloudlet, the user can offload the task request to the cloudlet server deployed on v_j with 1 hop, and the communication transmission delay can be counted as 0. Otherwise, the user u_i can only access the cloudlet placed with some other APs from v_j in a multi-hop manner. Meanwhile, the communication access delay of AP cannot be ignored. Given the delay tolerance Dt_r of a user u_r, the average delay of the wireless APs transmitting user task requests to cloudlet does not exceed the given user delay tolerance Dt_r. The key notations are listed in Table 1.

2.2. **Problem definition.** In the MEC architecture, the minimization of the number of heterogeneous cloudlet deployments in the WMAN is defined as follows. In the WMAN

TABLE 1. Symbol notations

Symbols	Definition
$G = \{V \cup S \cup U, E\}$	APs set, potential cloudlet locations set and the mobile users set.
$m = \|V\|$, $\in = \|E\|$, $n = \|U\|$	The number of APs in V, the number of links in E, and the number of users in U.
v_i	The i-th AP in set V.
u_i	The i-th user in set U.
R_j	User request collection of AP v_j.
q_{ik}	Contact probability between user u_i and AP v_k.
Z_{ik}	The number of tasks that AP v_k receives from user u_i.
du_i	The resource demand of user u_i.
$d(e)$	Link delay between APs.
Dt_r	Delay tolerance of user u_r.
d_{ij}	Transmission delay between AP v_j and v_i.
D_{ij}	The latency of user u_i offloading task request to a cloudlet located at v_j.
D_{avg}	The average delay of APs transmitting user requests.
$Z(v_k)$	The number of task requests that AP v_k received.
r_k	Resource capacity of cloudlet server f_k.
φ_{ij}	A binary variable indicating whether the task request user u_i is offloaded to a cloudlet located at v_j.
γ_{ij}	A binary variable indicating whether the cloudlet server f_k is deployed at AP v_i.

system $G = \{V \cup S \cup U, E\}$. Given a user's tolerable average time delay Dt_r for transmitting user requests by a wireless AP, k cloudlet servers $\{f_1, f_2, \ldots, f_k\}$ with k resource capacities $\{r_1, r_2, \ldots, r_k\}$, which need to be deployed to k different potential locations, $f_i \neq f_j$, $r_i \neq r_j$. The number of task requests from user u_i received by wireless AP v_k is denoted as $Z_{ik} = R_k * q_{ik}$, where $q_{ik} \in (0, 1)$. Therefore, the number of task requests from all users received by wireless AP v_k is denoted as $Z(v_k)$, the resource demand of $u_i \in U$ is described as du_i, $du_i \leq r_1 \leq r_2 \leq \cdots \leq r_k$. For each link (v_i, v_j) in E, define the latency of transmitting a user request between two APs v_i and v_j as the shortest path value between the two wireless APs, d_{ij} represents the latency of transmitting user requests between v_i and a cloudlet located at AP v_j, d_{ij}: $E \mapsto \mathbb{R} \geq 0$ [36]. Therefore, the problem is defined as while ensuring the users' QoS, k cloudlet servers with different resource capacities need to be placed in k potential deployment locations, aiming to minimize the number of k heterogeneous cloudlet deployments while the average delay for the wireless APs transmitting user requests to cloudlet cannot exceed the users' tolerable average delay Dt_r, that is, according to the different user densities and user resource demands in the WMAN, place k cloudlet servers with different capacities in k potential locations in S, aiming at minimizing the number of k heterogeneous cloudlet servers. Previous studies have demonstrated that the problem of cloudlet placement and user association in WMANs is an NP-hard problem [35,37]. Similarly, the problem of minimizing the number of heterogeneous cloudlet deployments is also an NP-hard problem.

2.3. **Problem formulation.** The problem of heterogeneous cloudlet deployment in the WMAN can be formulated as an ILP, and then a set of variables is defined to describe the heterogeneous cloudlet placement problem. When minimizing the problem of the number

of heterogeneous cloudlet deployments, for the QoS of users, given a user tolerable delay Dt_r, the average transmission delay of wireless APs transmitting the user task requests to cloudlet placed with the wireless AP shall not exceed the users' tolerable delay Dt_r. When selecting a wireless AP to place the cloudlet, several problems should be considered, including how to choose a potential location in S for cloudlet deployment, how many cloudlet servers should be deployed for each wireless AP, which cloudlet available for mobile users' task requests should be offloaded to and processing, the resource demands of mobile users and the resource capacity of cloudlet. Consequently, a binary variable $\gamma_{ij} \in \{0,1\}$, $1 \le i \le k$, $1 \le j \le |S|$ is introduced to represent whether cloudlet is deployed to location wireless AP $v_i \in S$, where $\gamma_{ij} = 1$ if cloudlet f_i is deployed to $v_i \in S$, $\gamma_{ij} = 0$ otherwise. A binary variable $\varphi_{ij} \in \{0,1\}$, $1 \le i \le n$, $1 \le j \le |S|$ is introduced to indicate whether the user task requests are offloaded to cloudlet for processing, where $\varphi_{ij} = 1$ if user task requests are offloaded to cloudlet placed with v_j, $\varphi_{ij} = 0$ otherwise. N_{ik} is denoted as the number of task requests from user $u_i \in U$ received by $v_k \in V$. R_k indicates the task request collection at v_k. q_{ik} is used to represent the contact probability between users u_i and v_k. Therefore, $0 \le Z(v_k) \le R_k$, $Z(v_k)$ is expressed as the number of task requests from all users received by $v_k \in V$. Define the latency of transmitting a user request between two APs v_i and v_j as the shortest path value between the two wireless APs, d_{ij} represents the latency of transmitting user requests between v_i and a cloudlet located at $v_j \in S$, D_{ij} is denoted as the latency of user u_i offloading task request to a cloudlet located at v_j, which is calculated as Equation (4).

$$D_{ij} = \sum_{k \in V} N_{ik} * d_{kj} * \varphi_{ij} \quad \forall j \in S \quad \forall i \in U \tag{4}$$

The average delay of all wireless APs V transmitting users' task requests to cloudlets is described as Equation (5).

$$Dt_r = \frac{\sum_{j=1}^{|S|} \sum_{i=1}^{|U|} D_{ij}}{\sum_{k=1}^{|V|} N(v_k)} \tag{5}$$

While not violating the QoS of users, the minimization of the number of heterogeneous cloudlet server deployments is formulated as follows.

$$\text{Obj: } \min k \tag{6}$$

$$\text{s.t. } \varphi_{ij} = \begin{cases} 1, & \text{if task request of user } u_i \text{ is offloaded to a cloudlet} \\ & \text{located at AP } v_j \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$\gamma_{ij} = \begin{cases} 1, & \text{if cloudlet server } f_i \text{ is deployed to AP } v_j \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\sum_{j=1}^{|S|} \gamma_{ij} = 1 \quad \forall i \in k \tag{9}$$

$$\sum_{i=1}^{|F|} \gamma_{ij} = 1 \quad \forall j \in S \tag{10}$$

$$\sum_{i \in U} \varphi_{ij} * q_{ik} * du_i \le \sum_{k \in F} r_k * \gamma_{jk} \quad \forall j \in S \quad \forall i \in V \tag{11}$$

$$Z(v_k) \le R_k \tag{12}$$

$$Z(v_k) = \sum_{i \in U} Z_{ik} \quad \forall k \in V \tag{13}$$

$$\frac{Z_{ik}}{Z(v_k)} \leq \sum_{i=1}^{|k|} \gamma_{ij} \quad \forall j \in S \tag{14}$$

$$\frac{\sum_{j=1}^{|S|} \sum_{i=1}^{|U|} D_{ij}}{\sum_{i=1}^{|V|} Z(v_k)} \leq Dt_r \tag{15}$$

Without loss of generality, where constraint (9) ensures that a cloudlet server can be only deployed to one potential deployment location S, and constraint (10) ensures that each potential access point should be deployed with at most one cloudlet server selected from the set F, which means that we do not require all cloudlet servers in the candidate set to be deployed in the MEC environment. The wireless AP where the cloudlet server is deployed can be associated with multiple users for service access, and different users have different task requests and therefore different resource requirements, in order to avoid overloading of cloudlet server resources, therefore, constraint (11) ensures that the total resource demand of mobile users shall not exceed the resource capacity provided by the cloudlet server, and constraint (12) ensures that some mobile users can be connected to $v_i \in V$ and wireless AP $v_i \in V$ can transmit mobile user requests to the cloudlet placed with potential location $v_j \in S$, and constraint (13) ensures that all user requests at $v_i \in V$ can be offloaded to cloudlets. Constraint (14) indicates that some user task requests at $v_i \in V$ are assigned to potential location $v_j \in S$ for processing at any time, which means that each user's requests at the wireless AP $v_i \in V$ are assigned to the potential location $v_j \in S$, and there must be a cloudlet server deployed on $v_j \in S$. To ensure users' QoS, depending on the cloudlet relationship, the expected service access delay of any user shall not exceed the tolerable service access delay; therefore, constraint (15) ensures that the average transmission latency of all APs transmitting users' task requests D_{avg} does not exceed the given users' tolerable average access latency Dt_r.

3. **Heterogeneous Cloudlet Deployment Algorithm Design.** In MEC, the minimization of the number of heterogeneous cloudlet deployments based on the WMAN scenario is an NP-hard problem, and it is formulated as an ILP problem. Due to the poor scalability of the ILP problem, it is important to design a low-complexity heuristic algorithm to solve the ILP problem [35,37]. However, in [35], the heuristic algorithm selects the optimal AP for the cloudlet deployment according to the order of the transmission delay between wireless APs, and ignores the mobility of users and the heterogeneity of cloudlet servers. In [37], it ignores the average latency of APs transmitting user requests and the resource demands of user tasks. Therefore, IMHA is proposed, which calculates the average access delay by contacting probability q_{ij} between user u_i and AP v_j and the shortest data transmission delay between AP v_j and v_i calculated by Floyd algorithm, which can balance the workload of wireless APs, user density and transmission cost between wireless AP points. Then, sort the APs according to $T(v_j)$ and determine optimal APs for cloudlet deployment, which is calculated as Equation (16).

$$T(v_j) = \frac{\sum_{i=1}^{m} q_{ij} * d_{ij}}{m} \quad \forall j \in V \tag{16}$$

where m is the number of APs connected to AP v_j. And IMHA is divided into two sub-problems, including the selection problem of the cloudlet server and the deployment problem of the cloudlet server and the user and cloudlet association problem, that is, the average transmission latency of all APs transmitting users' task requests D_{avg} does not exceed the given users' tolerable average access latency Dt_r. Therefore, in order to facilitate the elaboration of the heuristic cloudlet deployment algorithm with improved

strategy, the IMHA is divided into the cloudlet deployment algorithm (Algorithm 1) and the heuristic algorithm for minimizing the number of cloudlet deployments (Algorithm 2).

3.1. **Improved cloudlet deployment algorithm.** In Algorithm 1, from Step1 to Step6, first, arrange the k cloudlet capacities in descending order. Then, for wireless AP v_i with user requests remaining, the average access delay $T(v_j)$ is calculated as Equation (16), and sort $T(v_j)$ in descending order. And the more details are from Step7 to Step10, and then allocated to the cloudlet f_i until the cloudlet capacity r_i is just used up, and calculate the delay $\sum_{i=1}^{|U|} D_{ij}$ of all user requests assigned to cloudlet f_i placed with wireless AP v_j in Step11. If the user meets the delay requirements, cloudlet f_i will be deployed to the location where the smallest delay D_{ij} of wireless AP transmitting user task requests to a cloudlet is placed with AP v_j.

Algorithm 1: Cloud Placement

Input: G, r_k, R_k, d_{ij}, du_i, Dt_r, q_{ij}
Output: D_{avg}, φ_{ij}
Step1: Find a subset of cloudlet servers that meet user resource demands
Step2: Sort each k cloudlet server in descending order of resource capacity
Step3: ***for*** $i \leftarrow 1$ to k ***do***
Step4: $v_i \leftarrow S$
Step5: ***for*** each potential location v_j ***do***
Step6: Sort $T(v_j)$ in descending order
Step7: Select the first r wireless APs after sorting until the resource demand requested by the user contacting the r wireless APs exactly exceeds the resource capacity r_k of f_k
Step8: Assign $r-1$ users to the cloudlet f_j deployed at wireless AP v_j
Step9: Sort the resources required for the r-th user request in ascending order
Step10: The user request of the r-th wireless AP is selected according to the descending order of the du_i and allocated to the cloudlet f_i until the cloudlet capacity is just used up
Step11: Calculate the total delay $\sum_{i=1}^{|U|} D_{ij}$ of all user requests assigned to cloudlet f_i placed with wireless AP v_j
Step12: ***end for***
Step13: Cloudlet f_i is deployed to potential location with smallest value D_{ij}
Step14: ***end for***
Step15: D_{avg}
Step16: return D_{avg}, φ_{ij}

3.2. **Improved algorithm for minimizing the number of heterogeneous cloudlet deployments.** Combined with Algorithm 1, the heuristic algorithm for minimizing the number of cloudlet with an improved strategy (IMHA) is shown in Algorithm 2. For each link (v_i, v_j) in E, Floyd algorithm is used to calculate the shortest path value between wireless APs in Step2. The average access delay $T(v_j)$ of each AP is sorted in descending order in Step4. k cloudlet servers $\{r_1, r_2, \ldots, r_k\}$, which are sorted in descending order, $r_1 \geq r_2 \geq r_3 \geq \cdots \geq r_k$. The greedy heuristic algorithm is adopted, when $k = 1$, select the cloudlet with the largest capacity all the time, so that k is incremented to n in steps of 1 until the number of cloudlet servers meets the resource demand required by the user's request and the D_{avg} does not exceed the given users' tolerable average access latency Dt_r. In each iteration, assume that cloudlet servers $f_1, f_2, \ldots, f_{k-1}$ have been deployed

Algorithm 2: IMHA

Input: G, r_k, $Z(v_k)$, d_{ij}, Dt_r, q_{ij}
Output: k', φ'_{ij}
Step1: $\varphi_{ij} = \{0\}$
Step2: The Floyd algorithm is used to calculate the shortest path between all wireless APs in network G
Step3: ***for*** each potential location v_j ***do***
Step4: Sort $T(v_j)$ in descending order
Step5: ***end for***
Step6: $\varphi_{ij} = \{0\}$
Step7: ***while*** $D_{avg} \geq Dt_r$ ***do***
$k' \leftarrow k + 1$
Step8: $\varphi'_{ij} = \{1\}$
Step9: $Dt_r =$ **Cloudlet Placement** $(G, R_i, k, S, r_1, r_2, \ldots, r_k, du_i)$
Step10: ***end while***
Step11: return $r_1, r_2, \ldots, r_k$, k', φ'_{ij}

to wireless APs $v_1, v_2, \ldots, v_{i-1}$. Therefore, wireless AP v_i has not deployed the cloudlet, and cloudlet f_k will be deployed to the location where the smallest delay D_{ij} of wireless AP transmitting user task requests to a cloudlet is placed with AP v_i.

Theorem 3.1. *Given a WMAN system $G = \{V \cup S \cup U, E\}$, the time complexity of IMHA is $O(kn^2 \log n + n^4)$, $1 \leq k \leq |S| \leq |V|$, $n = |V|$, $m = |E|$.*

Proof: Let $n_1 = |V| + |S|$ be the number of nodes in $G = \{V \cup S \cup U, E\}$. In order to construct the delay transmission matrix D_{ij}, the Floyd algorithm is used to calculate d_{ij}. Because the time complexity of calculating a single source node using the Floyd algorithm is $O(n_1^3)$, the shortest path between all wireless AP nodes as $O(n_1^4) = O(n^4)$. It can be seen from Algorithm 2 that if cloudlet servers $f_1, f_2, \ldots, f_{k-1}$ have been deployed to $v_1, v_2, \ldots, v_{i-1}$ respectively, in each iteration, the wireless node $v_i \in S$ must be found at the i-th iteration to place the cloudlet f_i, $|S| \leq |V| = n$, the time complexity $O(k \cdot |S| \cdot |V| \log |V|) = O(kn^2 \log n)$; therefore, the time complexity of IMHA is $O(kn^2 \log n + n^4)$.

4. **Experimental Analysis.** To be more practical, cloudlet placement is highly promising for the implementation of the MEC paradigm which has been widely used in many fields such as face recognition [9], healthcare [11], interactive gaming and augmented reality [15] to reduce the end-to-end latency perceived by the mobile users' task requests. Furthermore, a fundamental but important issue for ISP is how to plan the location and capacity of cloudlets in order to minimize the number of cloudlet servers to reduce the cost of ISP while ensuring users' QoS. Therefore, we evaluated the performance of the IMHA through four dimensions, including the number of mobile users, the number of cloudlet servers, the maximum resource demand of users, and the maximum resource capacity of cloudlet servers. We compared the existing research HA [35] with IMHA. MTD-HA represents the minimum transmission delay of the wireless APs transmitting user task requests solved by the HA, and MTD-IMHA represents the minimum transmission delay of the wireless AP transmitting user task requests solved by the IMHA; ATD-HA represents the average transmission delay of wireless APs transmitting user task requests solved by the HA, and ATD-IMHA represents the average transmission user request of wireless APs transmitting user task requests solved by the IMHA. The experimental settings are consistent with [35,37], and Barabasi-Albert model in the Networkx package in Python3.7 is

used to generate the underlying complex network topology. The link delay $d(e)$ is generated at [5ms, 50ms], the number of cloudlet servers is set to half of the number of wireless APs, $k = m/2$, and R_k is drawn from $[50, 500]$. Study the scalability of algorithms, by varying the number of users $user(n)$ from 5 to 50, and the number of cloudlet servers from 5 to 50. Then the resource capacity r_k of the cloudlet server f_k is drawn from $[10, 500]$, and the resource demand du_i requested by the user task is generated at $[1, 80]$.

4.1. **Effect of the number of users on the number of cloudlet servers.** In this chapter, we analyze the effect of the number of users on the number of cloudlet servers, the average transmission delay and the minimum transmission delay of all wireless APs transmitting user requests. The number of cloudlet servers is set to 10 and the number of APs is set to 20. And we set the number of users $user(n) \in [10, 50]$, since, both methods are 100% efficient at finding a feasible solution when $user(n) < 10$ and cannot find a feasible solution when $user(n) > 50$. The ATD-IMHA, MTD-IMHA, ATD-HA and MTD-HA are shown in detail in Figure 2. As the number of users increases, the ATD-IMHA and MTD-IMHA gradually decrease, and the ATD-HA and MTD-HA are basically unchanged. As the average transmission delay increases, the comparison of the number of cloudlet servers is shown in Figure 3, and when $r_k \in [250, 350]$, the number of users increases from 10 to 50. In order to verify the effectiveness of IMHA more accurately, one-way ANOVA is used to further calculate the variance between the ATD-IMHA and ATD-HA to analyze the effectiveness of IMHA, and the significance value α is set to 0.05, after calculating, P-value is 0.521523. Mathematically, H_0 hypothesis: $\mu_1 = \mu_2$, the alternative hypothesis H_1: $\mu_1 \neq \mu_2$. Since P-value: $0.521523 > \alpha$: 0.05, H_0 is accepted, which demonstrate that the difference between the average transmission delay of HA and IMHA is not big enough to be statistically significant.

Although the H_0 hypothesis is accepted, as shown in Table 2, due to the limitation of server resource capacity, as the number of users increases, the resource demand of user task requests increases. When $r_k \in [100, 200]$, $n = 30$, the number of cloudlet deployments solved by IMHA is 4, but there is no feasible solution when $n = 30$ through HA. Similarly,

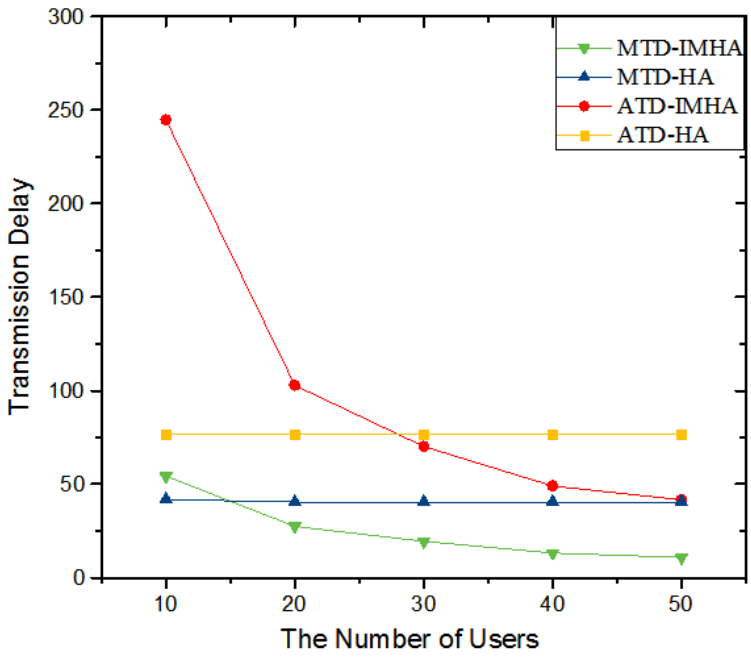

FIGURE 2. Effect of the number of users on the transmission delay

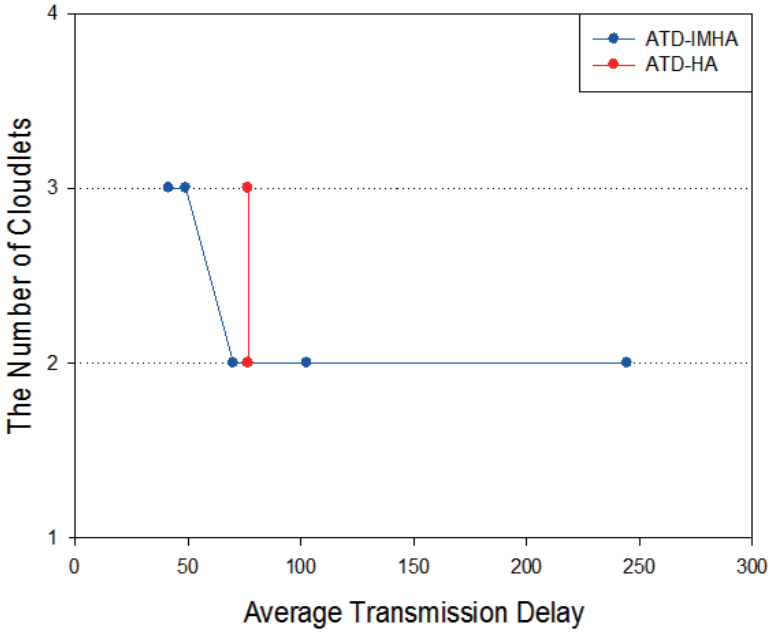

FIGURE 3. Effect of delay and the number of users on the number of cloudlets

TABLE 2. Effect of the number of users on the number of cloudlet servers

$user(n)$	$[100, 200]$		$[150, 250]$		$[200, 300]$		$[250, 350]$		$[350, 500]$	
	IMHA	HA	IMHA	HA	IMHA	HA	IMHA	HA	IMHA	HA
10	2	2	2	2	2	2	2	2	2	2
20	3	3	2	2	2	2	2	2	2	2
30	4	–	4	4	3	3	3	3	2	2
40	–	–	–	–	4	–	3	3	3	3
50	–	–	–	–	–	–	4	–	3	3

when $r_k \in [250, 350]$ and $n = 50$, the number of cloudlet deployments solved by IMHA is 4, but HA has no feasible solution. Only when the server resource capacity is large enough and the value is $[350, 500]$, HA can be found feasible solution. The above experimental results show that IMHA is more effective than HA.

4.2. **Effect of the number of servers on the number of candidate cloudlet servers.** In this chapter, we analyze in each iteration, given a maximum number of servers, how many available cloudlet servers are available, and which can meet the resource requirements requested by the user's task. The number of mobile users is set to 20, and the number of APs is set to 15. And we set the number of cloudlet servers $server(k) \in [4, 12]$, in a group of experiments, we noticed that if $server(k) < 4$, there is no feasible solution in most cases. Therefore, the case of more than 4 servers is only considered, and a feasible solution can be absolutely found when $server(k) > 12$. The average delay and minimum transmission delay of wireless APs transmitting user requests are shown in Figure 4. As the number of servers increases, MTD-IMHA and ATD-IMHA are lower than that of HA. When $server(k) \in [4, 12]$, $r_k \in [150, 250]$. The comparison of the number of cloudlet servers is shown in Figure 5. Similarly, the significance value α is set to 0.05, and after calculation, P-value is 0.258414. Mathematically, H_0 hypothesis: the alternative hypothesis H_1. Since P-value: $0.258414 > \alpha$: 0.05, H_0 is accepted, which indicates that the

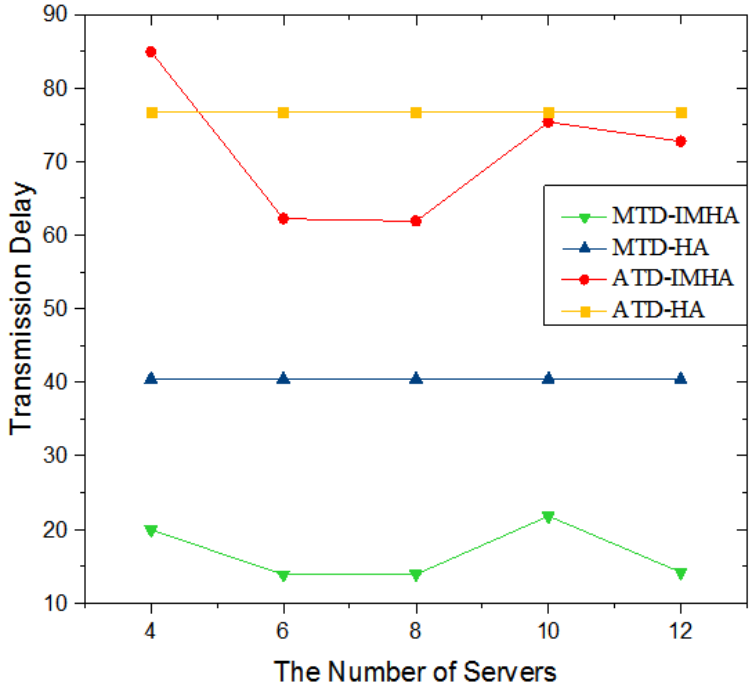

FIGURE 4. Effect of the number of servers on the transmission delay

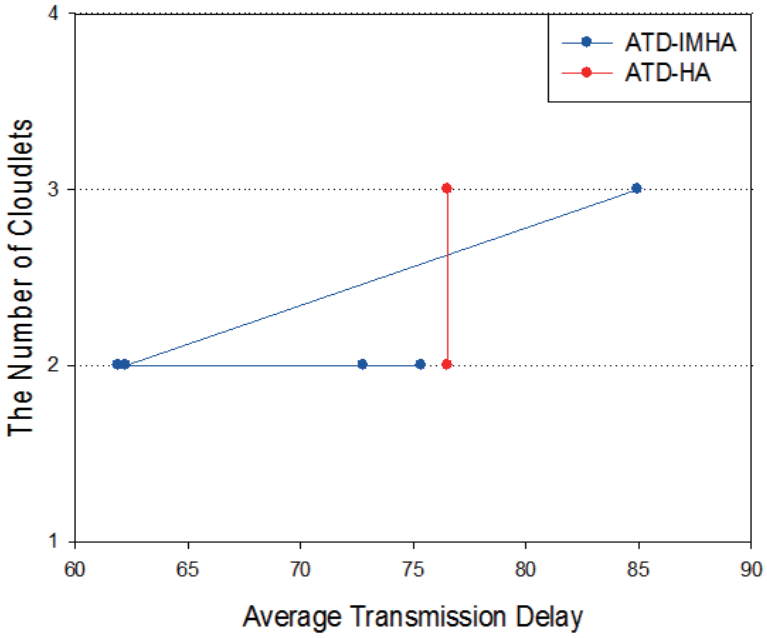

FIGURE 5. Effect of delay and the number of servers on the number of cloudlets

difference between the average transmission delay of HA and IMHA is not big enough to be statistically significant.

However, it can be seen from Table 3 that when $n = 8$, the resource capacity of the cloudlet server r_k is $[150, 250]$, and the number of cloudlet deployments solved by IMHA is 2, which can meet the resource requirements of users. However, only the number of cloudlets is 3 in HA, which can meet resource demand of users. Similarly, when $n = 10$, the number of cloudlets solved by IMHA is 2, the resource demand of users can be met. In HA algorithm, only the number of cloudlets is 3, which can meet the resource demand

TABLE 3. Effect of the number of servers on the number of candidate cloudlets

$server(k)$	$[100, 200]$		$[150, 250]$		$[200, 300]$		$[250, 350]$		$[350, 500]$	
	IMHA	HA	IMHA	HA	IMHA	HA	IMHA	HA	IMHA	HA
4	3	3	3	3	2	2	2	2	2	2
6	3	3	2	2	2	2	2	2	2	2
8	3	3	2	3	2	2	2	2	2	2
10	3	3	2	3	2	2	2	2	2	2
12	3	3	2	2	2	2	2	2	2	2

of users. With the increase of the number of cloudlets, the cloudlet server is already sufficient and stable to meet the needs of users, and there is no need to further include more candidate servers. This shows that IMHA has a high performance in solving the number of cloudlet deployments and transmission delays.

4.3. **Effect of the maximum resource demand of users on the number of cloudlet servers.** In this chapter, the number of users is set to 50, the number of wireless APs is set to 20, and the number of cloudlet servers is set to 10. The resource demand of the mobile user task request is generated in $[10, 80]$, after a group of experiments, both methods are absolutely efficient at finding a feasible solution when $du_i < 10$ and cannot find a feasible solution when $du_i > 80$. And the resource capacity of the server is drawn from $[10, 500]$. Meanwhile, as is shown in Figure 6, with the increase in the amount of resources required for user task requests, ATD-IMHA is lower than that of HA, and the MTD-IMHA is lower than that of HA. The comparison of the number of cloudlet servers is shown in Figure 7. Similarly, the significance value α is set to 0.05. P-value is calculated as 4.55191e-15. Therefore, H_0 hypothesis: $\mu_1 = \mu_2$, the alternative hypothesis H_1: $\mu_1 \neq \mu_2$. Since P-value $< \alpha$, H_0 is rejected, which demonstrates that the difference between the average transmission delay of HA and IMHA is big enough to be statistically significant.

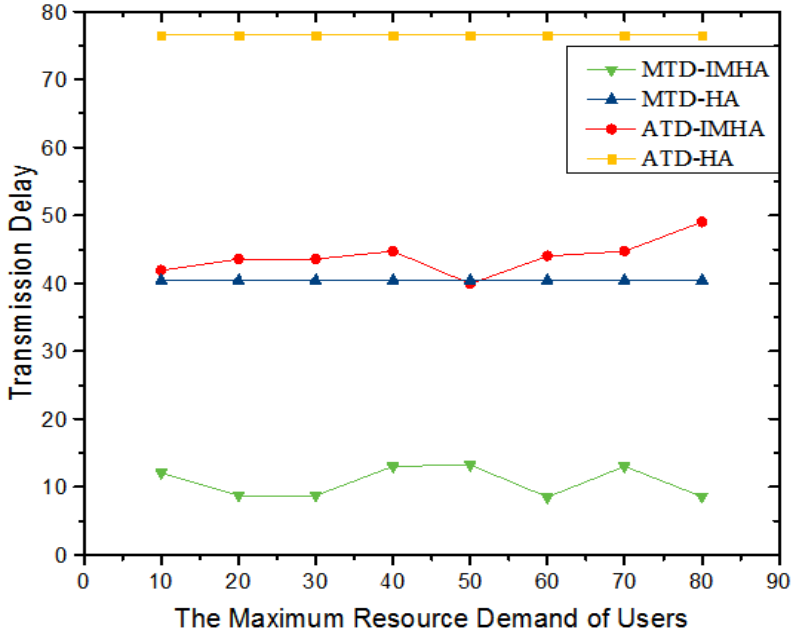

FIGURE 6. Effect of the maximum resource demand of users on the transmission delay

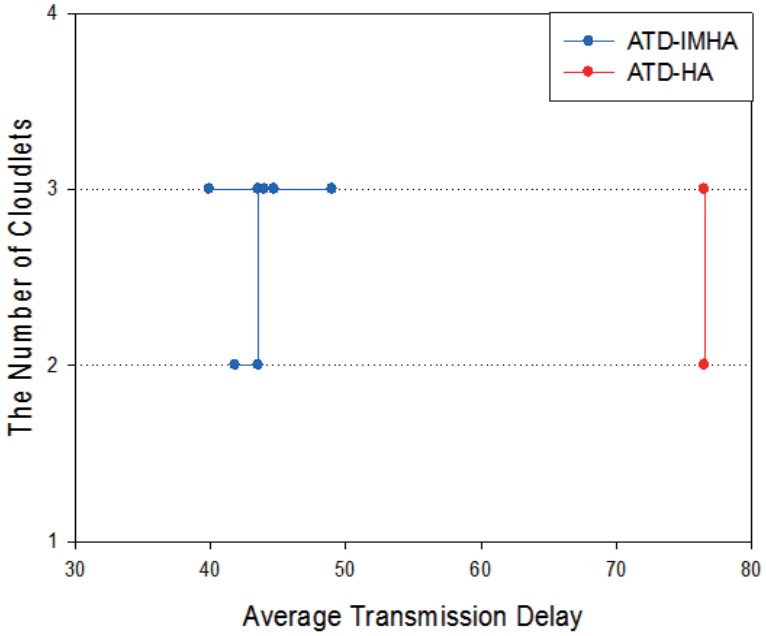

FIGURE 7. Effect of delay and the maximum resource demand of users on the number of cloudlets

At the same time, it can be seen from Table 4 that when the number of users is 50 and du_i is generated in $[50, 80]$, the number of cloudlet deployments solved by IMHA is 3, and HA has no feasible solution, only when du_i is drawn from $[40, 80]$, the number of users is 40, HA can find a feasible solution. Only when the number of users is 30 and du_i is generated in $[70, 80]$, HA can find a feasible solution.

TABLE 4. Effect of the maximum resource demand of users on the number of cloudlets

du_i	user = 50		user	HA
	IMHA	HA		
10	2	2	50	2
20	2	2	50	2
30	3	3	50	3
40	3	3	40	3
50	3	–	40	3
60	3	–	40	3
70	3	–	30	3
80	3	–	30	3

4.4. **Effect of the maximum resources capacity of servers on the number of cloudlet servers.** The number of users is set to 50, the number of wireless APs is set to 20, the number of cloudlet servers to 10, the minimum server resource capacity r_k is 10, and the maximum increases from 100 to 500. After a number of experimental tests, we noticed that if the maximum server resource capacity $r_k < 100$, there is no feasible solution in most cases. Therefore, a feasible solution can be absolutely found when the maximum server resource capacity $r_k > 500$. Because the resource capacity of the servers is affected by the resource demand of users. As is shown in Figure 8, as the maximum server resource capacity increases, MTD-IMHA and ATD-IMHA are lower than that of

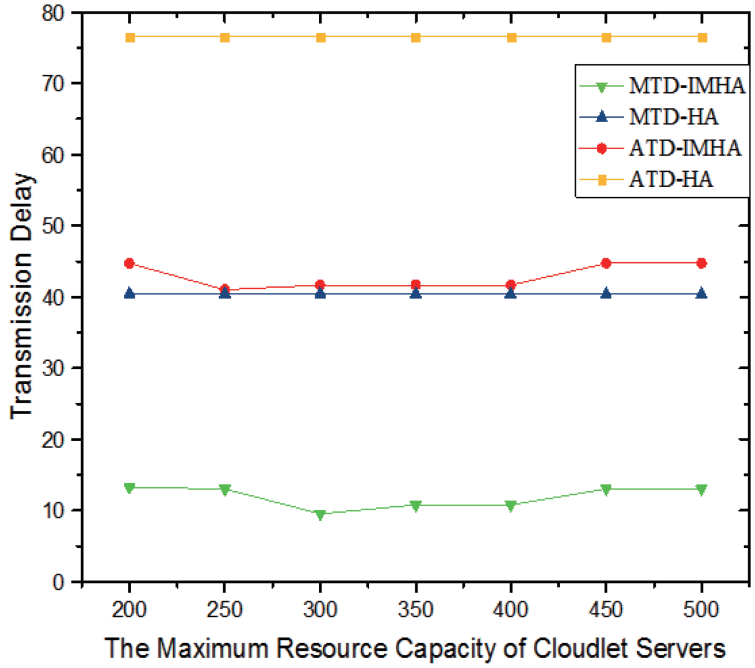

FIGURE 8. Effect of the maximum resource capacity of servers on the transmission delay

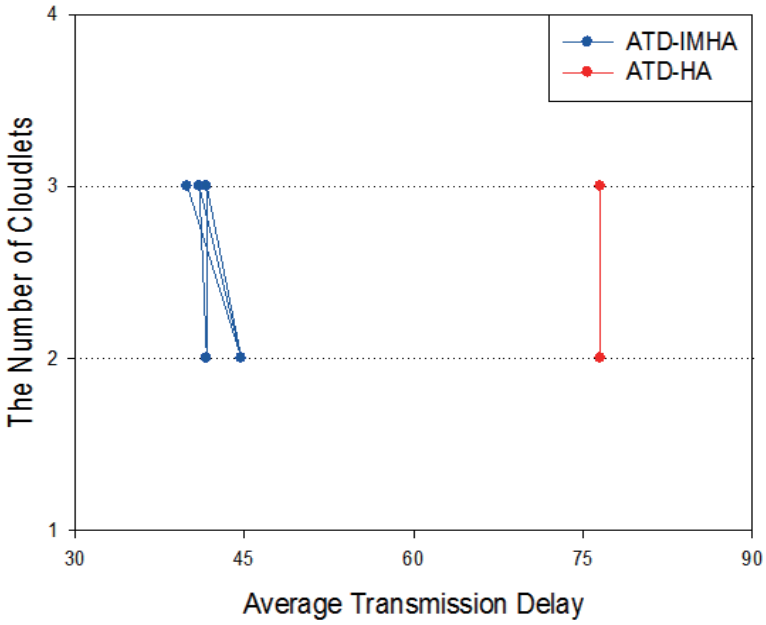

FIGURE 9. Effect of delay and the maximum resource capacity of servers on the number of cloudlets

HA. When $r_k \in [200, 500]$, the comparison of the number of cloudlets corresponds to the delay in Figure 9. α is set to 0.05, after calculation, P-value is 2.33147e-15. H_0 hypothesis: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$. Since P-value $< \alpha$, H_0 is rejected, which proves that the difference between the average transmission delay of HA and IMHA is big enough to be statistically significant.

Meanwhile, it can be seen from Table 5 that when the user's task requirement resource $du_i \in [1, 15]$ and the server resource capacity r_k is 250, the minimum number of cloudlet

TABLE 5. Effect of the maximum resources capacity of servers on the number of cloudlets

Server capacity (r_k)	du_i	IMHA	HA	du_i	HA
200	$[1, 10]$	3	3	$[1, 10]$	3
250	$[1, 15]$	2	3	$[1, 15]$	3
300	$[1, 20]$	3	3	$[1, 20]$	3
350	$[1, 25]$	2	2	$[1, 25]$	2
400	$[1, 30]$	3	3	$[1, 30]$	3
450	$[1, 35]$	2	–	$[1, 30]$	3
500	$[1, 35]$	2	–	$[1, 30]$	3

deployment solved by IMHA is 2; however, HA is 3. However, when $du_i \in [1, 35]$, IMHA can find a feasible solution, but HA has no solution. Similarly, only when $du_i \in [1, 35]$, can a feasible solution be found. These results shows that IMHA has high performance and efficiency in solving the number of cloudlet deployments and transmission delay.

5. **Conclusions.** MEC provides a method to reduce the latency of cloud services by building cloudlet infrastructure that are very close to the end users. In this article, we study the problem of heterogeneous cloudlet deployment and user mobility in WMAN, aiming at minimizing the number of heterogeneous cloudlet deployments while ensuring the QoS of users. We first considered the user dynamics, the heterogeneity of cloudlet capacity, and the average transmission delay for all wireless APs transmitting user requests to cloudlets, and designed a new model for minimizing the number of heterogeneous cloudlet deployments, and then this problem is formulated as an ILP problem, which is proved to be an NP-hard problem. Finally, the method of selecting an optimal wireless AP for cloudlet deployment is improved, and a heuristic algorithm with an improved strategy to minimize the number of heterogeneous cloudlet deployment (IMHA) is proposed to solve the problem. A large number of simulations show that the improved model and algorithm have higher performance and effectiveness. In the future, we will consider the cost and latency of cloudlet deployment as optimization goals, and study the combination of deployment and task offloading among multiple cloudlet provided by different edge service providers.

Acknowledgment. This research was supported by the National Natural Science Foundation of China (Grant No. 61762031), the Foundation of Guilin University of Technology (Grant No. GUTQDJJ2002018), and the Guangxi Universities Key Laboratory Director Fund of Embedded Technology and Intelligent Information Processing (Grant No. 2019-01-10).

REFERENCES

[1] X. Sun and N. Ansari, Latency aware workload offloading in the cloudlet network, *IEEE Communications Letters*, vol.21, no.7, pp.1481-1484, 2017.

[2] A. Jin, W. Song and W. Zhang, Auction-based resource allocation for sharing cloudlets in mobile cloud computing, *IEEE Transactions on Emerging Topics in Computing*, vol.6, no.1, pp.45-57, 2018.

[3] G. Wang, X. X. Huang and J. Zhang, Levitin-Polyak well-posedness in generalized equilibrium problems with functional constraints, *Pacific Journal of Optimization*, vol.6, no.2, pp.441-453, 2010.

[4] B. Qu and J. Zhao, Methods for solving generalized Nash equilibrium, *Journal of Computational and Applied Mathematics*, vol.2013, no.1, pp.111-134, 2013.

[5] E. Wong, M. P. I. Dias and L. Ruan, Predictive resource allocation for tactile Internet capable passive optical LANs, *IEEE Journal of Lightwave Technology*, vol.35, pp.2629-2641, 2017.

[6] X. Li, D. Li, J. Wan, C. Liu and M. Imran, Adaptive transmission optimization in SDN-based industrial Internet of Things with edge computing, *IEEE Internet of Things Journal*, vol.5, no.3, pp.1351-1360, 2018.
[7] M. Shen, B. Ma, L. Zhu, X. Du and K. Xu, Secure phrase search for intelligent processing of encrypted data in cloud-based IoT, *IEEE Internet of Things Journal*, vol.6, no.2, pp.1998-2008, 2019.
[8] Y. Wang, X. Sun and F. Meng, On the conditional and partial trade credit policy with capital constraints: A stackelberg model, *Applied Mathematical Modelling*, vol.40, pp.1-18, 2016.
[9] F. Zhang and M. M. Wang, Stochastic congestion game for load balancing in mobile-edge computing, *IEEE Internet of Things Journal*, vol.8, no.2, pp.778-790, 2021.
[10] T. V. Le and T. T. Huan, Computational intelligence towards trusted cloudlet based fog computing, *2020 5th International Conference on Green Technology and Sustainable Development (GTSD)*, Ho Chi Minh City, Vietnam, pp.141-147, 2020.
[11] S. Sengupta and S. S. Bhunia, Secure data management in cloudlet assisted IoT enabled e-health framework in smart city, *IEEE Sensors Journal*, vol.20, no.16, pp.9581-9588, 2020.
[12] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher and V. Younget, Mobile edge computing: A key technology towards 5G, Sophia Antipolis, France, *ETSI White Paper*, p.16, 2015.
[13] S. P. Ahuja and A. C. Rolli, Exploring the convergence of mobile computing with cloud computing, *Network & Communication Technologies*, vol.1, no.1, pp.1-97, 2012.
[14] T. Verbelen, P. Simoens, F. D. Turck and B. Dhoedt, Cloudlets: Bringing the cloud to the mobile user, *Proc. of the 3rd Workshop on Mobile Cloud Computing and Services*, Low Wood Bay, Lake District, UK, pp.29-36, 2012.
[15] S. Mondal, G. Das and E. Wong, Computation offloading in optical access cloudlet networks: A game-theoretic approach, *IEEE Communications Letters*, vol.22, no.8, pp.1564-1567, 2018.
[16] K. Jiang, H. Ni, R. Han and X. Wang, An improved multi-objective grey wolf optimizer for dependent task scheduling in edge computing, *International Journal of Innovative Computing, Information and Control*, vol.15, no.6, pp.2289-2304, 2019.
[17] D. Chatzopoulos, C. Bermejo, S. Kosta and P. Hui, Offloading computations to mobile devices and cloudlets via an upgraded NFC communication protocol, *IEEE Transactions on Mobile Computing*, vol.19, no.3, pp.640-653, 2020.
[18] Z. Wang, D. Zhao, M. Ni and C. Li, Collaborative mobile computation offloading to vehicle-based cloudlets, *IEEE Transactions on Vehicular Technology*, vol.70, no.1, pp.768-781, 2021.
[19] X. Zhu and M. C. Zhou, Multi-objective optimized cloudlet deployment and task offloading for mobile edge computing, *IEEE Internet of Things Journal*, 2021.
[20] F. Zhang, J. Ge, Z. Li, C. Wong and L. Kong, A load-aware resource allocation and task scheduling for the emerging cloudlet system, *Future Generation Computer Systems*, vol.87, no.10, pp.438-456, 2018.
[21] S. S. Chalapathi G., V. Chamola, C. K. Tham, S. Gurunarayanan and N. Ansari, An optimal delay aware task assignment scheme for wireless SDN networked edge cloudlets, *Future Generation Computer Systems*, vol.102, no.3, pp.862-875, 2019.
[22] X. H. Deng, J. Li, E. L. Liu and H. G. Zhang, Task allocation algorithm and optimization model on edge collaboration, *Journal of Systems Architecture*, vol.110, pp.101778-101791, 2020.
[23] J. Xu, Z. Hu and J. Zhou, Computing offloading and resource allocation algorithm based on game theory for IoT devices in mobile edge computing, *International Journal of Innovative Computing, Information and Control*, vol.16, no.6, pp.1895-1914, 2020.
[24] S. Yang, F. Li, M. Shen, X. Chen, X. Fu and Y. Wang, Cloudlet placement and task allocation in mobile edge computing, *IEEE Internet of Things Journal*, vol.6, no.3, pp.5853-5863, 2019.
[25] D. Rupanetti and H. Salamy, Task allocation, migration and scheduling for energy-efficient real-time multiprocessor architectures, *Journal of Systems Architecture*, vol.98, pp.17-26, 2019.
[26] S. Nithya, M. Sangeetha, K. N. A. Prethi, K. S. Sahoo, S. K. Panda and A. H. Gandomi, SDCF: A software-defined cyber foraging framework for cloudlet environment, *IEEE Transactions on Network and Service Management*, vol.17, no.4, pp.2423-2435, 2020.
[27] A. N. Asadi, M. A. Azgomi and M. R. Entezari, Analytical evaluation of resource allocation algorithms and process migration methods in virtualized systems, *Sustainable Computing: Informatics and Systems*, pp.100370-100385, 2020.
[28] X. Sun and N. Ansari, Adaptive avatar handoff in the cloudlet network, *IEEE Transactions on Cloud Computing*, vol.6, no.3, pp.664-676, 2019.
[29] Q. Fan and N. Ansari, On cost aware cloudlet placement for mobile edge computing, *IEEE/CAA Journal of Automatica Sinica*, vol.6, no.4, pp.926-937, 2019.

[30] C. Shen, S. Xue and S. C. Fu, ECPM: An energy-efficient cloudlet placement method in mobile cloud environment, *Journal on Wireless Communications and Networking*, vol.141, pp.1-10, 2019.
[31] L. Chen, J. G. Wu, G. Q. Zhou and L. Ma, Quick: QoS-guaranteed efficient cloudlet placement in wireless metropolitan area networks, *Journal of Supercomputing*, vol.74, no.8, pp.4037-4059, 2018.
[32] L. Zhao, W. Sun, Y. P. Shi and J. J. Liu, Optimal placement of cloudlets for access delay minimization in SDN-based Internet of Things networks, *IEEE Internet of Things Journal*, vol.5, no.2, pp.1334-1344, 2018.
[33] X. J. Guan, X. L. Wan, T. J. Wang and Y. F. Li, A long-term cost-oriented cloudlet planning method in wireless metropolitan area networks, *Electronics*, vol.8, no.11, pp.1-20, 2019.
[34] W. Wibisono, T. Ahmad, R. Anggoro and Rozita, A grid-based clustering with dynamic forwarding path for energy-efficient data gathering in wireless sensor network environments, *ICIC Express Letters, Part B: Applications*, vol.10, no.3, pp.185-193, 2019.
[35] Z. C. Xu, W. F. Liang, W. Z. Xu, M. Jia and G. Song, Efficient algorithms for capacitated cloudlet placements, *IEEE Transactions on Parallel and Distributed Systems*, vol.27, no.10, pp.2866-2880, 2015.
[36] L. Ma, J. Wu and L. Chen, DOTA: Delay bounded optimal cloudlet deployment and user association in WMANs, *2017 17th IEEE/ACM International Symposium on Cluster Cloud and Grid Computing (CCGRID)*, Madrid, Spain, pp.196-203, 2017.
[37] H. Yao, C. M. Bai, M. Z. Xiong, D. Z. Zeng and Z. J. Fu, Heterogeneous cloudlet deployment and user-cloudlet association toward cost effective fog computing, *Concurrency and Computation*, vol.29, no.16, pp.1-14, 2017.