Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH Open Access

An edge server deployment method based on optimal benefit and genetic algorithm

Hongfan Ye[1,2], Buqing Cao[1*], Jianxun Liu[1], Pei Li[1], Bing Tang[1] and Zhenlian Peng[1]

Abstract

With the speedy advancement and accelerated popularization of 5G networks, the provision and request of services through mobile smart terminals have become a hot topic in the development of mobile service computing. In this scenario, an efficient and reasonable edge server deployment solution can effectively reduce the deployment cost and communication latency of mobile smart terminals, while significantly improving investment efficiency and resource utilization. Focusing on the issue of edge server placement in mobile service computing environment, this paper proposes an edge server deployment method based on optimal benefit quantity and genetic algorithm. This method is firstly, based on a channel selection strategy for optimal communication impact benefits, it calculates the quantity of edge servers which can achieve optimal benefit. Then, the issue of edge server deployment is converted to a dual-objective optimization problem under three constraints to find the best locations to deploy edge servers, according to balancing the workload of edge servers and minimizing the communication delay among clients and edge servers. Finally, the genetic algorithm is utilized to iteratively optimize for finding the optimal resolution of edge server deployment. A series of experiments are performed on the Mobile Communication Base Station Data Set of Shanghai Telecom, and the experimental results verify that beneath the limit of the optimal benefit quantity of edge servers, the proposed method outperforms MIP, K-means, ESPHA, Top-K, and Random in terms of effectively reducing communication delays and balancing workloads.

Keywords Mobile service computing, Optimal benefit quantity, Load balancing, Communication delay, Edge server deployment

Introduction

With the rapid development of mobile Internet, mobile intelligent terminals fully penetrate social life and personate a mushrooming number of crucial roles in people's regular life [1]. With the advent of the 5G era [2], the concept of "premises on safety construction, guided by data aggregation and computing, and driven by an extensive smart applications" has promoted the development of smart cities [3], and it has also led to the explosive growth in mobile service computing [4]. Due to the limitation of mobile intelligent terminals' own computing resources, it is unable to process the growing requests in time for mobile service computing, which seriously affects the user experience of mobile users [5]. Therefore, edge computing is formally proposed and developed rapidly [6–8]. In edge computing environment, the functions of data processing are added at the edge close to data producers (that is, the uplink part of mobile service computing), and computing power and storage resources are distributed close to mobile intelligent terminals to supplement and optimize cloud computing. That is to say, the cloud is no longer responsible for all service requests in edge computing. The mobile intelligent terminal unloads the task to the edge server(ES) which is liable for base stations to

*Correspondence:
Buqing Cao
buqingcao@gmail.com
[1] The School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China
[2] School of Computer and Artificial Intelligence, Huaihua University, Huaihua, China

process through the mobile communication base stations and returns the processing results [9]. If the edge server cannot handle the task, it will be further unloaded to the cloud server [10, 11]. Therefore, in timely data analysis and intelligent handling of mobile service computing, edge computing has a broader application [12], and it is more effective and secure than simple cloud computing [13, 14] . In the domain of edge service computing, most research works currently focus on the service tasks of mobile smart terminals, such as service offloading [15–17], service migration [18–21], micro-service Mashup [22], and so on.

To achieve efficient edge service resource optimization, many researchers model it as an optimization problem to solve [23–26]. Furthermore, it is also a hot topic to introduce machine learning methods into edge computing [27–30]. However, few researchers have observed that the effective deployment of edge servers is a pre-requisite for large-scale mobile service computing applications [31]. An efficient and feasible edge server deployment has a significant impact on reducing deployment cost and communication delay, and improving efficiency and resource utilization [32, 33].

In view of this, this paper focuses on the issue of providing an efficient and feasible edge server deployment scheme in mobile service computing environment, and proposes an edge server deployment model named HE-GA, which is used to achieve edge server deployment with the lowest communication delay and the most load-balanced under the optimal benefit quantity limit. In this model, firstly, the volume benefit curve of deploying each edge server is calculated by probability modeling, and the optimal number of benefit deployment is obtained. Then, the load-balance and communication delay of the edge server are modeled. Finally, under the restriction of the optimal benefit quantity, it exploits genetic algorithm [34] to resolve the dual-objectives of balancing the load and reducing the communication delay, and gains an efficient and feasible deployment scheme of edge server. In summary, the main contributions of this paper are indicated as below:

- Aiming at the issue of edge server deployment, the HE-GA model is proposed to achieve the dual-objective optimization of communication delay and load balance under the premise of satisfying the optimal benefits quantity. The HE-GA model can be applied to edge server deployment in smart cities. As far as we know, few scholars have conducted in-depth research in this topic.
- Based on the HE-GA model, with the optimal benefit quantity limitation, the genetic algorithm is exploited to solve the dual-objective issue under multiple constraints to balance the load and reduce the communication delay, to obtain an efficient and feasible edge server deployment solution.
- On the basis of Mobile Communication Base Station Data Set of Shanghai Telecom, the comparative experiments and in-depth analysis are performed. The experimental results demonstrate that, under the premise of satisfying the optimal benefit quantity, the proposed approach comprehensively outperforms five typical deployment methods such as MIP, K-means, AK-means, Top-K, and Random according to reducing the communication delay and balancing the workload.

The organization of this paper is arranged as: Introduction section briefly presents the research background and the main work of this paper; Related work section introduces the related work; HE-GA method section describes the provided HE-GA model exhaustively; Experiment and analysis section gives the experimental results and analysis. Eventually, Conclusion and prospect section summarizes this paper and expects its future work.

Related work

With the large-scale commercial application of the 5G network, service computing is penetrating social life in an all-around way at an unprecedented speed [35]. Meanwhile, mobile intelligent terminals are widely used in our daily life, resulting in an unprecedented demand for ultra-low latency, powerful computing, and storage capacity [36, 37]. Therefore, distributed deployment of computing power and resources close to customers has become a trend [38]. In this context, mobile service computing can be regarded as an effective solution to promote the high-quality development of smart city construction [4, 6]. In the field of mobile service computing, there are many high-quality researches in the domestic and international, most of which focus on service offloading, service migration, server deployment and other aspects [7, 8, 39].

However, every service computing task can only be performed after the successful placement of edge servers [40]. So effective placement of edge servers is becoming a significant challenge in the field of mobile service computing [41], and this has led to a wealth of research results. For example, Yin et al. [42] proposed a decision-making sustain framework called Tentacle to deploy edge servers. Tentacle exploits dominant position of the increasingly elastic layout of edge servers to find appropriate unforeseen edge locations to optimize the total property and price of edge infrastructure, thus significantly improving efficiency and reducing the price of edge configuration. Cui et al. [43] modeled the problem of joint user coverage and edge server placement for network robustness, then proposed an optimal method based on integer programming. Wang et al. [32] formulated the edge server deposition issue of intelligent cities in mobile edge computing environments as a multi-objective constrained optimization issue. Then, to counterpoise the workload of edge server and minimize the access delay among mobile intelligent terminals and edge servers, they use MIP to gain the optimal response. Guo Feiyan [44] raised a mobile edge server placement medium called ESPHA based on an improved heuristic algorithm by combining K-means algorithm and AG algorithm to minimize access delay and load difference. Kasi et al. [39] formulated the edge server deposition issue as a multi-objective constrained optimization issue and then applied Genetic Algorithm and fractional search algorithm to obtain the optimal edge server deposition scheme. To provide low-latency and low-cost services in VANET networks, Zhang et al. [40] proposed a joint optimization approach for the deployment locations of edge servers and service coverage in urban areas. To address the limitations of service-intensive deployment in 6G environments, Cong et al. [9] presented a mobile resource sharing framework utilizing mobile edge servers. This framework enables large-scale edge resource sharing in the context of the IoT environment. Cui et al. [43] designed a redundant service deployment model in a heterogeneous edge environment to achieve high-quality QoS. The model targets the redundant deployment of services for different applications and utilizes a priority-based genetic algorithm to obtain an optimized solution. To obtain a more optimal resource allocation solution, Asghari et al. [41] refined the cellular mobile network into smaller regions and used the CRO algorithm to optimize resource allocation in each region. Fan et al. [38] proposed a novel joint resource management scheme based on the time-slot system that leverages end-to-cloud collaboration, which effectively ameliorates the performance of DNN inference services in industrial IoT applications.The aforementioned works have optimized the deployment of edge servers from different perspectives and achieved good results. However, these works have not considered how to minimize the deployment cost of edge servers while obtaining maximum performance, i.e., achieving the deployment of edge servers with the lowest latency and balanced load under the constraint of optimal benefits. This would effectively balance the conflicting interests of operators and end users.

Inspired by above researches and the best communication collision benefit designed by Ji et al. [45], this paper proposes an edge server deployment method named HE-GA to obtain the lowest delay and the most load balance under the optimal benefit quantity limit, which is used to achieve high-quality distributed deployment of edge servers. The experimental results on the data set of mobile communication base station of Shanghai Telecom show that under the optimal benefit quantity limitation, the HE-GA method effectively counterpoises the load of the edge server and reduces the communication delay between edge servers and the base stations.

HE-GA method

This section includes four parts: Motivation section presents the research motivation of this paper, then lists the main symbols used below; Edge server deployment model section details the edge server deployment model; Finally, genetic algorithm is exploited to solve the issue in Genetic algorithm section.

Motivation

The edge server deployment is composed of cloud center layer, edge node layer, and terminal node layer in a typical mobile service computing system. And the main components of it include:

- **Cloud Service Center:** it is still consistent with current cloud computing center. In mobile service computing system, it possesses the most powerful computing power and storage resources in the system. It provides all services for end users. The cloud service center communicates with all edge nodes. It is responsible for processing tasks with huge computing power demand and storing all results. At the same time, the cloud service center carries out distributed policy distribution management for the tasks of edge mobile service computing system.
- **Edge Node:** it is mainly composed of edge server which has certain computing power and storage resources. It provides services to terminal users within its ability through reasonable deployment. Every edge server is liable for handling terminal user requests in a certain area through the base station. All edge servers unite to achieve the full coverage of the intelligent city at edge layer. Simultaneously, all edge servers are linked to cloud service center. They become the connection hub between the end users, mobile communication base stations, and the cloud service center. The number and manner of deployment of edge servers are also the core of this paper's research.
- **Terminal Node:** it mainly refers to mobile intelligent terminal. It has extremely limited computing power and storage resources. Its main task is to establish contact with the edge server through base stations, transmit the mobile intelligent terminal's service requisition to edge nodes or cloud service center, and present the final result to the user.

For the sake of the easement of characterization, the main insignias used below are enumerated here, as Table 1 shows.In mobile service computing environment, the edge server, communication base station, and mobile

Table 1 Symbol system

Symbol	Implication
G	Mobile Service Computing network
V	The node of Mobile Service Computing
E	Connection during base station and edge server of Mobile Service Computing network
B	The set of mobile correspondence base station
n	Quantity of mobile communication base station
S	Set of edge server
K	Quantity of edge server
E_s	Set of base stations responsible for every edge server
t_b	The workload of base station b, where $b \in B$
t_s	The workload successfully assigned to edge server
t_c	Edge servers deploy additional fixed load
T_s	The workload of the edge server s, where $s \in S$
T_s^{opt}	Optimal number of cost-effective deployments of edge servers
ρ	The optimal benefits of edge servers
l_b	Situation of base station b
l_s	Situation of edge server s
d	Communication delay among edge server and basestation
K_s	Optimal deployment amount of edge servers
p	Probability of a base station choosing an edge server
P_s	Probability of successful allocation of all base stations

intelligent terminal can be modeled as a network, as Fig. 1 shows. The network can be defined as an undirected graph $G = (V, E)$, when $V = B \cup S(B$ represents base stations, S represents edge servers), and E indicates the communication link among base stations and edge servers. No connection between any two base stations, but they communicate through their respective edge servers. The communication links among mobile intelligent terminals and base stations are ignored because the issue is not the focus of this paper. Each base station and edge server have a fixed service area. The mobile intelligent terminal communicates with the edge server through the communication base station responsible for its own area and obtains the corresponding services. The number of mobile intelligent terminal services received by each edge server is regarded as its actual workload. The distance between edge servers and mobile communication base stations is regarded as the communication delay. In the following, we utilize t_b to represent the workload of base station b, l_b to represent the location of base station b, and l_s to indicate the possible locations of edge server s, where $b \in B, s \in S$.

Some key problems need to be identified. For example, How to calculate the optimal benefit quantity $K(K = 1, 2, 3...)$ of the edge server, under the limitation of the optimal benefit quantity, how to select $K(K = 1, 2, 3...)$ locations from the existing base station locations to deploy edge servers. In addition, how to achieve simultaneous optimization of communication delay and load equilibrium of edge server deployment under the premise of satisfying the optimal benefit quantity (dual-objective optimization)? It should be noted that in realistic scenarios, edge servers may be homogeneous or heterogeneous in each procurement and deployment, i.e., the computational and storage resources of edge servers may not be uniform. Therefore, it is indeed necessary to consider the deployment of edge servers in homogeneous and heterogeneous conditions separately in realistic scenarios. As the deployment of homogeneous edge servers still has certain practical significance, and this paper focuses on how to reasonably allocate existing mobile communication base stations to edge servers, the impact of the computing and storage resources of edge servers themselves on resource scheduling is not the main focus of this paper. Therefore, for the sake of modeling simplicity, this paper assumes that all edge servers have consistent computing power and storage resources, thus simplifying the application scenario. To sum up, based on realistic scenarios, the following assumptions are made in this paper: 1. Each edge server has the same delimited computing power and storage resources to process the service request of mobile intelligent terminal, and each base station will directly communicate with its own edge server to offload the tasks and return the result. 2. No base station is shared between every edge server, and the sum of base stations responsible for edge servers is B. 3. The total workload assigned to each edge server does not exceed a predefined maximum value. Subsequently, the optimization objectives can be represented as: 1. The workload can be allocated to each edge server in a balanced way. 2. Minimize the correspondence delay between edge servers and base stations. Take edge server S_2 in Fig. 1 as an example. Suppose there are four mobile communication base stations that communicate with the edge server S_2 and request respective services. The requested load by each base station is 4 for b_1, 5 for b_2, 8 for b_3, and 9 for b_4, respectively. Thus, the total load to be handled by the edge server S_2 is the sum of the loads requested by the mobile communication base stations $b_1 - b_4$, i.e., $T_{s_2} = 26$. It is necessary to identify a suitable location l_{s_2} to deploy the edge server S_2 to minimize the total communication delay d_{s_2} between the edge server S_2 and the four base stations. At the same time, it is necessary to balance the workload of the S_2 and other edge servers. That is, when deploying edge servers S_1,S_3, it is necessary to minimize the gap between the total load T_{s_1} ,T_{s_2}, T_{s_3} handled by edge servers S_1,S_2,S_3.

Edge server deployment model

Setting of constraints and calculation of optimal benefit quantity

In above mobile edge computing network, suppose deployment locations of the edge servers as

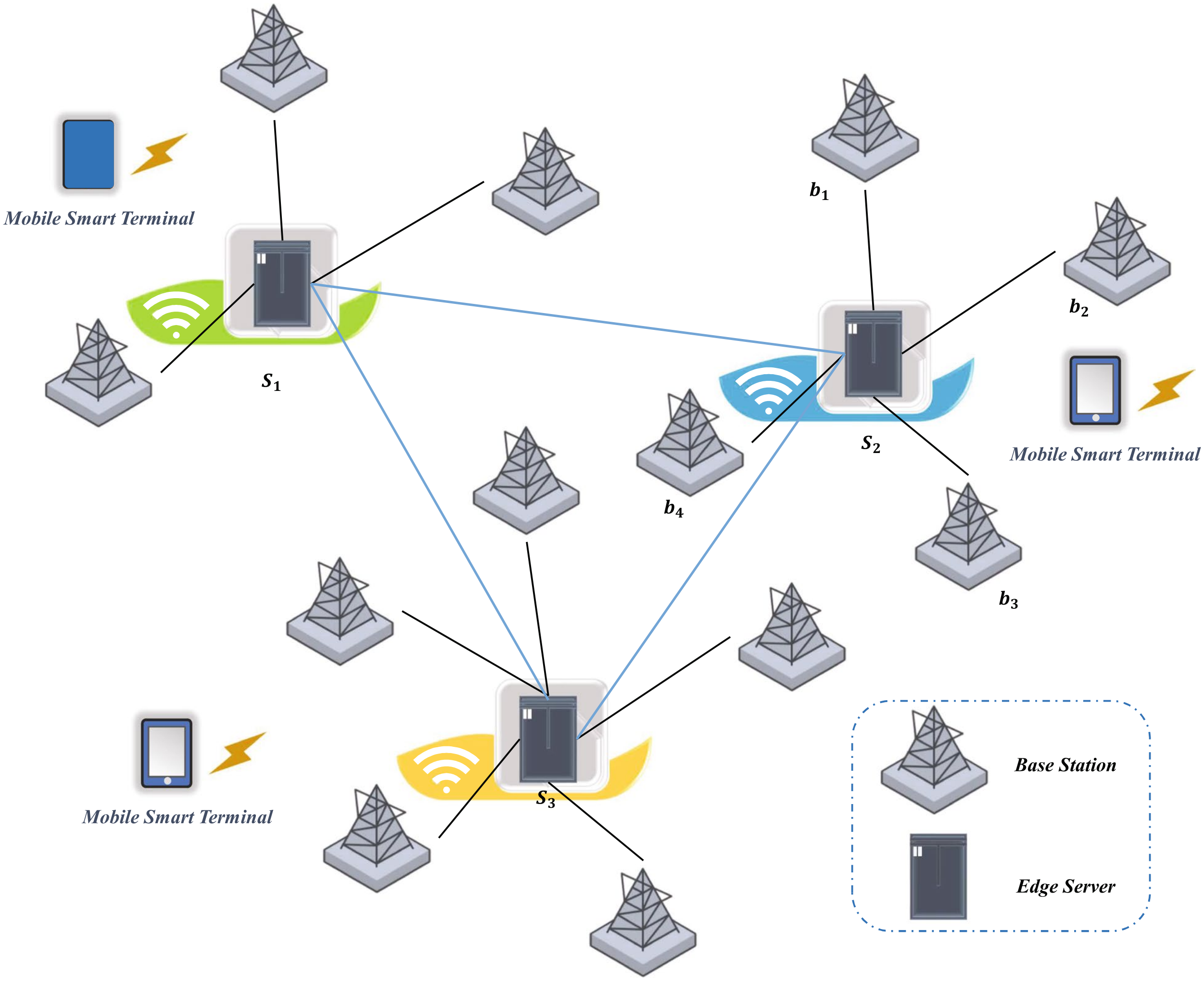

Fig. 1 Communication scene of mobile edge computing network

$l = l_1, l_2, l_3, \ldots, l_K$, the edge server is represented as $S = s_1, s_2, s_3, \ldots, s_K$, and the three constraints of the edge server deployment problem are formalized as follows:

- **Constraint 1:** Any two edge servers do not share any base station directly, and the total amount of base stations that all edge servers are responsible for justly is mobile communication base station set B.
- **Constraint 2:** For the edge server, any service request of the mobile intelligent terminal that establishes contact with it through its responsible mobile communication base station must be processed.
- **Constraint 3:** To ensure the highest input-output benefit of mobile edge servers, the amount of deployed edge servers is less than the optimal benefit quantity, and the total load on each edge server does not exceed a predefined maximum value.

The following Formula (1) - Formula (7) will deduce for constraint 3 in detail.

Lemma 1 *If the edge server deployment benefit rate is* $\rho = t_s/(\sum T_s + t_c)$*, the number of edge server deployment with the highest benefits number is* $T_s^{opt} = arg\ max\ \rho$.

Proof

Ji et al. [45] raised a channel selection method derived from the optimal communication collision benefit in 2014. Inspired by this, we design a work benefit function for edge server deployment problems as follows:

As all edge servers in the application scenario of this paper are assumed to be homogeneous, it is convenient for calculation to assume that the hardware and deployment cost of each edge server are uniform. Thus, the deployment cost of a single edge server can be defined as the unit cost of deploying an edge server. Therefore, it can be assumed that the number of deployed edge servers is directly proportional to the deployment cost.

Assuming the total amount of base stations to be distributed is n, the total load of the i-th base station is t_i, the load will be allotted to the edge server, and E_s is the assemblage of base stations responsible for each edge server, then the total load T_s of every edge server is:

$$T_s = \sum_{b \in E_s} t_b \tag{1}$$

In the edge server deployment problem, every base station will be distributed to an edge server to be responsible for communication. There are two types of base station allocation states:

1 Unallocated status: the target base station has not been allocated yet.
2 Successful allocation status: the target base station has been successfully allocated to an edge server.

As all edge servers in the application scenarios considered in this paper are assumed to be homogeneous, the problem can be analogized to the following problem: there is a batch of available storage bins. The capacity of each storage box is equal and meets the requirements. We try to evenly distribute n balls into K bins under the premise of using as few bins as possible and maximize the use of each storage bin as possible. In this scenario, mobile communication base stations can be analogized to balls, and edge servers can be analogized to storage boxes. The probability p of each storage box being selected is equal. Therefore, the possibility of selecting any one of the K servers of one mobile base station is as below:

$$p = 1/K \tag{2}$$

Since there are n mobile communication base stations to be assigned, an edge server selects a base station while not selecting other base stations. At this time, the probability of a base station being allocated is $p(1-p)^{(n-1)}$, and the process is repeated for n times. Then the probability P_s that all base stations have been assigned states on an edge server is:

$$P_s = n * p(1-p)^{(n-1)} \tag{3}$$

Then, according to the above obtained probability, the number of successfully assigned base stations on K edge servers in a single complete assignment is:

$$n_s = K * P_s \tag{4}$$

Thus, the successfully distributed workload t_s to edge servers is:

$$t_s = \sum_{i=1}^{n_s} t_i * P_s \tag{5}$$

All other additional costs, such as hardware cost and communication loss, are converted into additional fixed load, which is recorded as t_c. The deployment benefits rate ρ of edge server is as follows:

$$\rho = \frac{t_s}{\sum T_s + t_c} \tag{6}$$

Therefore, the number of deployments with the highest benefits for edge server deployment is:

$$T_s^{opt} = arg\ max\ \rho \tag{7}$$

□

Multi-objective optimization modeling

In the deployment process of an edge server, it is expected to achieve the lowest communication delay and the most balanced load simultaneously. Therefore, the deployment issue of edge servers can be converted to a multi-objective optimization problem under multiple constraints. That is to say, under the limits described in Setting of constraints and calculation of optimal benefit quantity section, the workload of any two edge servers is minimized, and the communication delay between edge servers and mobile communication base stations is also minimized [32]. They can be formulated as the below formulas:

$$T(l) = MinMax(T_i - T_j), \forall i,j \in K \tag{8}$$

$$D(l) = MinMax\ d(l_i, l_j), \forall i,j \in K \tag{9}$$

Let L be total possible edge server deployment scheme, then $l \in L$ is an edge server deployment scheme including K edge server deployment locations. $T(l)$ and $D(l)$ respectively represent the minimum load balancing of edge servers and the minimum access delay between edge servers and mobile communication base stations under scheme l. Let E_s is the set of base stations served by each edge server. At the same time, the constraints are formalized as follows:

- All edge servers will be deployed, and no base station is shared between every edge server, and total base stations will be allotted to edge servers, i.e.,

$$E_i \cap Ej = \emptyset \tag{10}$$

$$\bigcup_{s \in S} = B \tag{11}$$

- The deployment location of the edge server will be opted from the deployment locations of the mobile communication base stations, and the location will be shared with the base stations. Through the mobile communication base station responsible for the area, any mobile intelligent terminal will correspond with edge servers responsible for base stations and request services. The edge server will process the service request from the user and return the result (as Formula (1) shows).
- To ensure the highest input-output benefit of mobile edge servers, the number of deployed edge servers must be conformed to the limit of optimal benefit quantity of deployed edge servers, and the total load on each edge server does not exceed a predefined maximum value, i.e.,

$$T_s \leq T_s^{opt}, T_s \leq argmax2t_s \tag{12}$$

Generally speaking, the longer the physical/communication distance between two communication entities, the longer the signal propagation time/the routing transit time, resulting in a higher overall communication delay. Therefore, we define the distance between the mobile communication base station l_b and the edge server at l_s as the communication delay $d(l_b, l_s)$, and the distance unit km is exploited as the basic unit for measuring communication delay.

In a word, we take all the mobile communication base station locations as candidates and find the best K locations to deploy K edge servers on premise of meeting three constraints and simultaneously achieving two optimization objectives.

Accordingly, the edge server deployment issue can be transformed to a single-objective optimization under multiple constraints with optimal solutions in a weighted form. The concrete process is described as below.

By integrating Formula (1) - Formula (12), and summarizing all assumptions and limitations in the paper, the edge server deployment problem in mobile service computing network can be defined by the following formula:

$$\begin{cases} find\ l = (l_{s_1}, l_{s_2}, ..., l_{s_k})^T \\ which\ minT(l), minD(l) \\ subject\ to\ E_i \cap E_j = \emptyset, \bigcup_{s \in S} E_s = B, T_s \leq T_s^{opt}, T_s \leq argmax2t_s \\ where\ T_s = \sum_{b \in E_s} t_b \end{cases} \tag{13}$$

Where l is a m-dimensional decision variable, which represents the possible deployment position of the K edge server, $minT(l)$ and $minD(l)$ are the optimization objective functions of l under three constraints. Therefore, Formula (1) - Formula (12) are converted to a dual-objective optimization issue with three constraints recorded as Issue 1.

Therefore, to gain Pareto optimal solution or the weak Pareto optimal solution of Issue 1, we utilize weighting medium to convert the dual-objective optimization issue under the three constraints expressed by Formula (13) to the single-objective optimization issue under the three constraints.

Suppose the weighting coefficients as w_1 and w_2, where w$w_1 \geq 0$, $w_2 \geq 0$, and $w_1 + w_2 = 1$. In this way, Issue 1 is converted to a single-objective optimization issue under three constraints called as Issue 2:

$$\begin{cases} find\ l = (l_{s_1}, l_{s_2}, ..., l_{s_k})^T \\ which\ min(w_1 * T(l) + w_2 * D(l)) \\ subject\ to\ E_i \cap E_j = \emptyset, \bigcup_{s \in S} E_s = B, T_s \leq T_s^{opt}, T_s \leq argmax2t_s \\ where\ T_s = \sum_{b \in E_s} t_b, w_1, w_2 \geq 0\ and\ w_1 + w_2 = 1 \end{cases} \tag{14}$$

The following will prove the edge server deployment issue is a NP-hard problem.

Lemma 2 *The deployment issue of edge servers in mobile service computing network $G = (V, E)$ is a NP-hard problem.*

Proof

We turn the K-median issue into an edge server deployment problem. That is to say, given an integrated graph $G' = (V', E')$, we measure its K-median. Therefore, we first construct a mobile service computing network $G = (V, E)$ from G', when $V = V'$, $E = V'$. The next problem is to choose a suitable location from G to deploy K edge servers. Aiming at this issue, we imitate it as a single-objective optimization with three constraints in G' in polynomial time complexity. The K-median problem of G' is justly the optimal solution for deploying edge servers *in G. Since K-median issue is a NP-hard problem [46], the edge server deployment issue in this paper is also a NP-hard problem.*

The following parts will prove that the optimal solution of Issue 2 is justly the optimal solution of Issue 1.

Theorem 1 *The solution of Issue* 2 *is the weak Pareto optimal solution of Issue* 1.

Proof

If $l' \in L$ is the solution of Issue 2, then invite $l' \in L$ not be the weak Pareto optimal solution of Issue 1. It must be a solution $l \in L$ making $T(l) \leq T(l')$, $D(l) \leq D(l')$. In the light of the assumption, when the weighting moduli $w_1, w_2 \geq 0$, then $w_1 * T(l) + w_2 * D(l) \leq w_1 * T(l') + w_2 * D(l')$. These conflicts the issue of supposing l' is the solution of Issue 2. Therefore, the hypothesis is not true, and l' is the weak Pareto optimal solution of Issue 1.

Theorem 2 *If the weighting moduli $w_1, w_2 \geq 0$, the solution of Issue* 2 *is the Pareto optimal solution.*

Proof

If $l'' \in L$ is the solution of Issue 2 with correct weighting coefficient, then let $l'' \in L$ not be the Pareto optimal solution. There must be a solution $l \in L$ making $T(l) \leq T(l')$, $D(l) \leq D(l')$, and $T(l) \leq T(l'')$, $D(l) \leq D(l'')$. When the weighting coefficient $w_1, w_2 \geq 0$, then $w_1 * T(l) + w_2 * D(l) \leq w_1 * T(l'') + w_2 * D(l'')$. These conflicts the problem of supposing that l'' is the solution of Issue 2. Thus, the assumption is not true, l'' is the Pareto optimal solution.

Genetic algorithm

In this section, GA [34] is exploited to solve the single-objective problem under the three constraints. It's a random entire search and optimization medium. In this process, the optimal search space is automatically obtained and guided. The search direction is adaptively adjusted to acquire the optimal solution simultaneously. The general MIP method to solve this issue is iterative operation, but the generalized iterative medium is prone to mire in the local minimum pitfall, then emerge in an endless loop making the iteration impossible. As a global optimization algorithm, the GA well conquers the weakness. Moreover, the minimized communication delay and workload balance in this paper belong to a discrete optimization problem. The GA has a good application in this field, while other heuristic algorithms, such as particle swarm optimization, are more suitable for solving and optimizing some continuous problems. Therefore, GA is utilized to optimize the deployment of edge servers in mobile service computing environment, then some basic concepts and settings of GA are defined as follows:

1 **Population:** Different biological individuals combine to form different groups, such a group is called a population. The edge server set is respectively regarded as the edge server population.
2 **Individual:** Single organisms that make up a population. Each single edge server in the edge server set is justly an individual.
3 **Gene:** A DNA fragment containing biological genetic information, that is, the individual genetic characteristics of the edge server. We utilize the transformed binary code of location information in the edge server as genes.
4 **Phenotype:** The algorithm forms the external performance of individual according to the genetic information, that is, the external performance of individual is formed according to the genes of the edge server.
5 **Adaptability:** The organisms that survive and reproduce in competition are better adapted to the environment. So, the adaptability is the evolutionary direction of the population. In the edge server deployment problem, the adaptability function is justly the final objective optimization function.
6 **Heredity:** In the process of reproduction, the genes will be replicated and crossed normally, and they will mutate with a low probability. In the deployment of edge server, each generation is a set of solutions.
7 **Natural Selection:** Individuals with high adaptability to their living environment in competition have more opportunities to participate in reproduction, and their offspring will be more and more. We utilize Stochastic Tournament to perform natural selection. That is to say, a pair of individuals are selected by roulette for each time, we let them to compete with each other, leaving individuals with high adaptability, then iterating until full edge servers are selected.
8 **Evolution:** In this process, the adaptability of organisms to the external environment (objective function) is utilized as the criterion, and the traits of organisms are constantly improved and evolved which defined as an optimization process in the issue of edge server deployment.

```
Input: The location and the load of each mobile communication base station
The quantity of edge servers needs to deploy $K$
The quantity of optimal benefit deployment $T_s^{opt}$
Output: Edge servers deployment locations
The load standard deviation and communication delay
// $P_c$, The probability of crossover events in genetic time
// $P_m$, The probability of variation events in genetic time
// $M$, Generating the size of edge server population $Pop$
// $EG$, Algebra of evolutionary iterations
// $T_f$, If an individual fitness function in the edge server population generated by evolution
  exceeds$T_f$, stop evolution
// Convert the location information of the edge server into binary code as individual genes
BEGIN
repeat
    Randomly initialize the gene sequences of all individuals in $Pop$ of all edge server
      populations and combine them into chromosomes;
    Calculate the fitness $F(i)$ of all the edge server population $Pop$ (Formula $(14)$) and sort;
    Initialize the empty edge server population $newPop$;
    repeat
        In the light of the fitness value, exploit the selection algorithm to select $2$ populations
          from the edge server population $Pop$;
        while random(0,1)< $P_c$ do
            Perform a crossover operation with probability $P_c$ on the selected gene fragments
              with the same population position number;
        end
        while random(0,1)< $P_m$ do
            Perform reverse mutation with probability $P_m$ on the binary chromosomes of the
              selected two edge server populations;
        end
    until If $K$ offsprings are created;
    Calculate the adaptability of the edge server population;
    Replacing mobile base station population $Pop$ with edge server population $newPop$;
until The adaptability score of all chromosomes exceed $T_f$, or the reproduction number
  exceeds $EG$;
END
```

Algorithm 1 HE-GA

Algorithm 1 shows the process of optimizing and solving the edge server deployment problem. It can be seen from Algorithm 1 that the edge server deployment problem to be solved by the GA is simulated as a biological evolution process, and the optimal solution of the issue is sought through continuous iterative evolution. The concrete process of it is as follows:

1 Randomly initialize the binary gene sequence of the edge server.
2 Take the objective function (formula 14) as the fitness of the individual.
3 Assemble the genes of each population edge server as chromosomes.
4 Screen the individuals of organisms based on their adaptability.
5 The selected offspring is constantly reproduced through genetic operators, i.e., duplication, crossover, mutation, to complete natural selection and evolution.
6 When all chromosomes' adaptability scores exceed the threshold or reach the preset number of iterations, stop the loop to obtain the last generation population, i.e., edge server deployment location.

Since the individual adaptability in the final population obtained by genetic algorithm is relatively high, the minimum value of the multimodal function has a higher probability to exist in this population. If we take the matter having supreme adaptability in the population as the solution of the edge server deployment issue, the solution has a higher probability to be the optimal solution to the edge server deployment issue.

Experiment and analysis

Data set and experimental settings

By the speedy advancement and accelerated penetration of 5G networks, the construction requirement of mobile communication base stations is growing rapidly. Its siting

Table 2 Shanghai mobile communication base station data set - statistics of some base stations

Base Station ID	Quantity of Terminals	Load /min
12	354	24958
400	1242	61972
40	1824	2571744
664	151	190730
328	108	140960
6023	261	14026

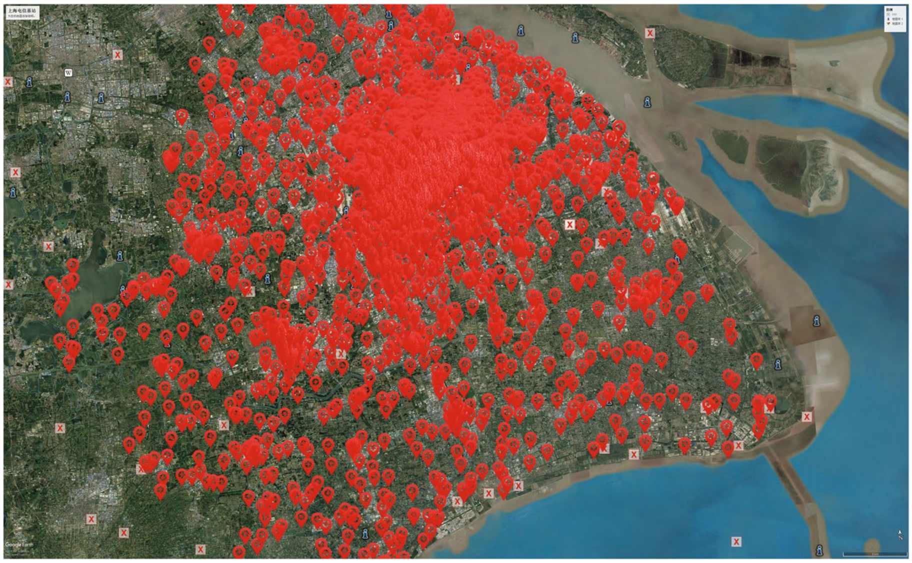

Fig. 2 Shanghai telecom mobile communication base station data set-distribution map of mobile communication base station

and deployment are becoming increasingly important. The selection principles of urban area site are: 1. Set up in the central area of the target load distribution as much as possible. 2. It should consider geography of an urban area and meet the requirements of network cellular topology. 3. The network optimization method is mainly used to form good complementarity with other surrounding stations, to balance the distribution of load among various sectors, especially to refrain the impact of cross-area coverage caused by excessive base station distribution on network performance. The non-urban site selection focuses on suburban factories, development zones, villages and towns, highways, large factories and mines, and these areas should be taken as key coverage objectives. It must have a definite object in view to improve the efficiency of base station construction and cover places with real load requirement. As everyone knows, Shanghai is one of Chinese supreme cities with a large population base and high density, so the data set in this area can better reflect the real situation of mobile service computing network environment. Therefore, to effectively evaluate the proposed HE-GA model, we perform the experiments on the real data set of Shanghai Telecom base station[1]. This data set includes the locations of 3233 Shanghai mobile communication base stations and the service request time information of the mobile intelligent terminals they serve. Through data analysis and cleaning, we find that there are some invalid data in the data set.

There are about 3000 valid base station information in 3233 base station data. There is no computing resource data such as CPU and memory in the data set. It only includes the initiate time and complete time of the access of mobile intelligent terminal. Therefore, we regard the application service time of the mobile intelligent terminal as the workload of edge servers and the distance between base stations calculated from the base station location information as the communication delay. Table 2 shows the statistical information of some base stations in the data set of Shanghai mobile communication base station.

The Fig. 2 illustrates the assignment of 3233 mobile communication base stations in the data set of Shanghai Telecom mobile communication base station. Each red landmark in the Fig. 2 represents the deployment position of a mobile communication base station, and we can see that the overall distribution of it is uneven. Among them, the upper part of Fig. 2 is the distribution of base stations in key urban areas such as schools, commercial

[1] http://sguangwang.com/TelecomDataset.html.

centers and railway stations, and its coverage density is higher; the left and the lower part of Fig. 2 are mainly the distribution of base stations in other non-urban areas, and its coverage density is relatively low. In this context, the deployment of base stations should balance the interests of operators and the needs of mobile intelligent terminal users, while considering the load of edge servers and mobile communication delays under the limitation of optimal benefits, which is the motivation of this research. The experiment is performed in the programming environment of Python 3.6, and the GA package in the Scikit-opt swarm intelligence optimization algorithm library is utilized to achieve the optimization solution of edge server deployment. Through multiple adjustments, the optimal parameters of the GA are identified, i.e., the initial population=200, the number of iterations=500, and the mutation rate=0.5.

Evaluation metric

Load balancing

The standard deviation of load between edge servers is utilized as the evaluation metric of load balancing. If K edge servers must to be deployed in the mobile communication base station group, the standard deviation of load between edge servers is defined as:

$$WB = \sqrt{\frac{\sum_{i=1}^{K}(T_i - \overline{T})^2}{K}} \tag{15}$$

Where T_i is the workload of the i-th edge server, $\overline{T}$ is the average workload of all edge servers. From Formula (15), we can see that the smaller the standard deviation of load between edge servers, the more balanced its load.

Communication delay

Another metric is the communication delay between edge servers and mobile communication base station. Due to the lack of communication delay data in the data set, we utilize the average distance between the edge server and the mobile communication base station to represent the communication delay.

Deployment rate

The deployment rate of edge server is defined as $ER = K/n$, when K is the quantity of edge servers deployed, then n is the number of mobile communication base stations.

Baseline methods

- **MIP** [32]**:** The mixed integer programming (MIP) method utilizes binary decision variables to indicate whether to allocate mobile communication base stations to edge servers, and the location of each base station is a candidate for edge server deployment. For each candidate value, a tag is given to consider both distance and workload to determine whether it is suitable for deploying edge servers.
- **ESPHA** [44]**:** Firstly, it combines the K-means algorithm and the ant colony algorithm to introduce a pheromone feedback mechanism in the deployment of the edge server. And then, it sets the taboo table in the ant colony algorithm to accelerate the algorithms' convergence. At length, the ameliorated heuristic algorithm is utilized to gain the optimal deployment scheme of edge servers.
- **SA:** It simulates the annealing process of solid materials in a physical scenario. Then, it takes a relatively high initial temperature as the starting state and undergoes a temperature parameter reduction process. At the same time, it employs the joint probability mutation characteristic to randomly search for the global optimal solution of the objective function.
- **K-means:** First, it randomly assigns K cluster centers in the mobile communication base station cluster, and divides the base stations to be classified into each cluster according to the nearest neighbor principle. Then it calculates the centroid of each cluster iteratively according to average medium until the shift distance of the cluster core is less than the threshold. The determined centroid situation is the deployment position of the edge server.
- **Top-K:** The top K mobile communication base stations having the top workload are picked as the deployment locations of the edge servers, and the base stations are allocated to the neighboring edge servers according to the principle of proximity.
- **Random:** K edge servers are randomly placed in the mobile communication base station cluster, and the base stations are apportioned to adjacent edge servers according to the principle of proximity.

Results and analysis

Tables 3, 4, 5, 6, 7, and 8, and Figs. 3, 4, 5, and 6 respectively present all the experimental results and their corresponding variation curves. From experimental results, it can be seen that the performance of HE-GA is superior than that of other methods in communication delay and workload balance. By summarizing the experimental results, we believe that the following two issues need to be considered for the deployment of edge servers.

- The relationship between cost and benefit should be considered during the investment in infrastructure

Table 3 Communication delay v.s. the quantity of base stations under optimal benefit limit(a) (*km*)

The number of base stations	300	600	900	1200	1500
HE-GA	3.9502	5.9305	6.5572	6.6250	5.4217
MIP	4.1954	7.4327	6.8487	6.8657	5.7346
ESPHA	4.2046	7.3610	6.6979	6.7329	5.6321
SA	6.3521	7.9482	6.9611	7.1527	6.2057
K-means	4.2573	7.0673	6.4209	6.7311	5.8539
Top-K	6.2476	7.7591	9.2102	12.7001	10.8571
Random	6.4843	11.0005	10.4465	12.2637	9.6427

Table 4 Communication delay v.s. the quantity of base stations under optimal benefit limit(b) (*km*)

The number of base stations	1800	2100	2400	2700	3000
HE-GA	5.5255	6.5455	5.4572	5.4357	5.3217
MIP	5.7244	6.9515	5.6781	5.6014	5.5346
ESPHA	5.6400	6.6672	5.5741	5.5120	5.4879
SA	6.2470	7.3143	6.3521	6.4163	7.1763
K-means	5.8417	6.6875	5.7186	5.6749	5.5741
Top-K	9.0193	9.7727	9.0147	8.7964	8.4104
Random	8.8271	11.5267	8.5687	8.2146	8.0973

Table 5 Workload balancing v.s. the quantity of base stations under optimal benefit limit(a)($*10^6$*min*)

The number of base stations	300	600	900	1200	1500
HE-GA	1.8255	5.7435	7.0572	8.2374	7.5217
MIP	1.9244	6.9741	7.4578	8.5471	8.4346
ESPHA	2.2162	6.1678	7.5127	8.3439	7.7398
SA	2.0086	7.1951	8.3902	9.9573	10.4094
K-means	2.1568	8.4070	10.4571	13.2346	13.1050
Top-K	1.1945	1.5696	1.4604	1.5138	1.4520
Random	2.2442	2.4327	2.6711	2.5928	2.4019

Table 6 Workload balancing v.s. the quantity of base stations under optimal benefit limit(b)($*10^6$*min*)

The number of base stations	1800	2100	2400	2700	3000
HE-GA	5.5255	6.5455	5.4572	5.4357	5.3217
MIP	5.7244	6.9515	5.6781	5.6014	5.5346
ESPHA	5.6400	6.6672	5.5741	5.5120	5.4879
SA	6.2470	7.3143	6.3521	6.4163	7.1763
K-means	5.8417	6.6875	5.7186	5.6749	5.5741
Top-K	9.0193	9.7727	9.0147	8.7964	8.4104
Random	8.8271	11.5267	8.5687	8.2146	8.0973

Table 7 Communication delay v.s. the quantity of edge servers of fixed mobile communication base stations (*km*)

Number of Edge Servers	100	200	300	400	500
HE-GA	8.9874	7.6874	5.3217	4.2174	3.4871
MIP	10.2587	8.5471	5.5346	4.8574	4.1789
ESPHA	9.8741	8.0665	5.4879	4.4674	3.8967
SA	10.6269	9.3239	7.1763	5.4677	4.9711
K-means	11.2413	7.5741	5.5864	5.0147	4.4893
Top-K	14.2146	10.6082	8.4104	6.0736	5.6747
Random	15.6712	13.8441	8.0973	6.4476	5.7410

Table 8 Workload balancing v.s. the quantity of edge servers of fixed mobile communication base stations ($*10^6$*min*)

Number of Edge Servers	100	200	300	400	500
HE-GA	16.3853	13.2547	7.7698	6.6250	5.0214
MIP	18.3746	15.3247	8.3210	6.8657	5.5698
ESPHA	17.4789	14.4226	7.9874	6.7329	5.4593
SA	19.3569	16.2525	9.5626	7.2916	6.5046
K-means	25.7821	20.2214	16.8412	17.6984	14.2373
Top-K	8.3698	4.2843	1.5782	1.2473	1.0247
Random	2.8741	2.3659	1.9874	1.4789	1.2589

such as mobile communication base stations, edge servers, and cloud service centers. Although the more infrastructure investment, the better technical indicators, but the corresponding cost of investment is higher. If the investment is too high, the rationality of the investment is debatable compared to its return. So, we specifically design the calculation method of the optimal benefit quantity. This is one of the research motivations in this paper. And none of the baseline approaches in this paper, including ESPHA, considers the optimal benefit quantity of edge server deployment.

- Optimal benefit, communication delay between edge server and mobile communication base station, load balancing value of edge server all need to compromise with each other, to gain the optimal deployment scheme, but difficult to achieve the optimum for all three simultaneously.

Analysis of the results of different quantity of base stations under the optimal benefit quantity limit

We utilize the data of 3000 effective base stations in the Shanghai Telecom mobile communication base station

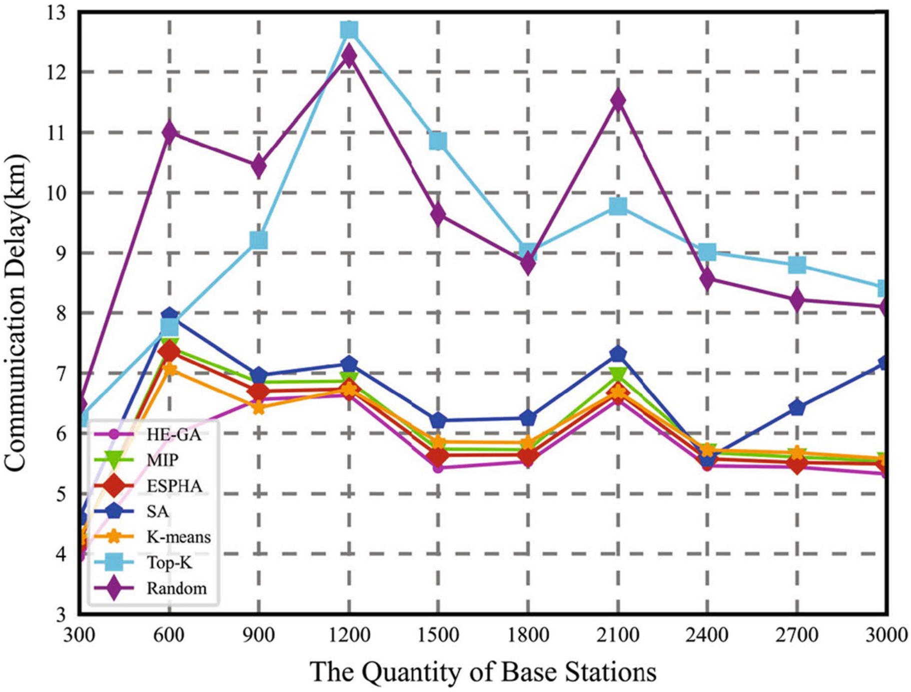

Fig. 3 Communication delay v.s. the quantity of base stations under optimal benefit limit

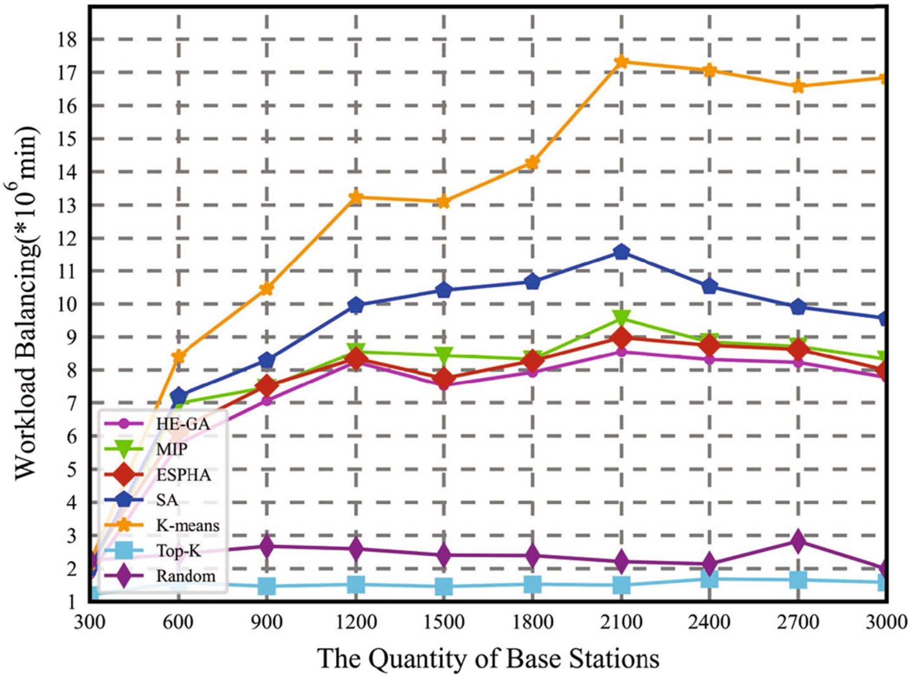

Fig. 4 Workload balancing v.s. the quantity of base stations under optimal benefit limit

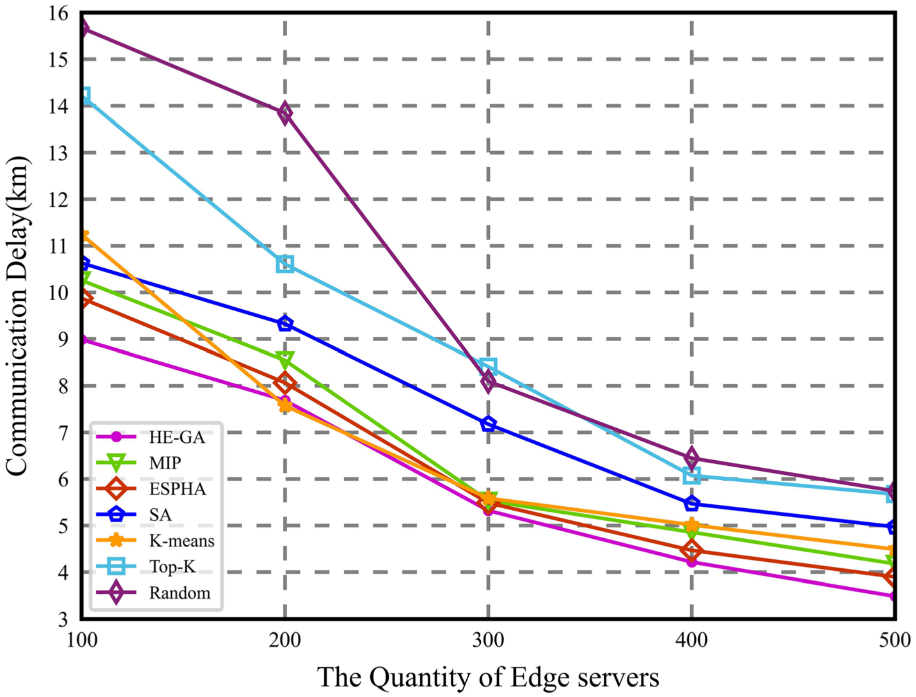

Fig. 5 Communication delay v.s. the quantity of edge servers of fixed mobile communication base stations

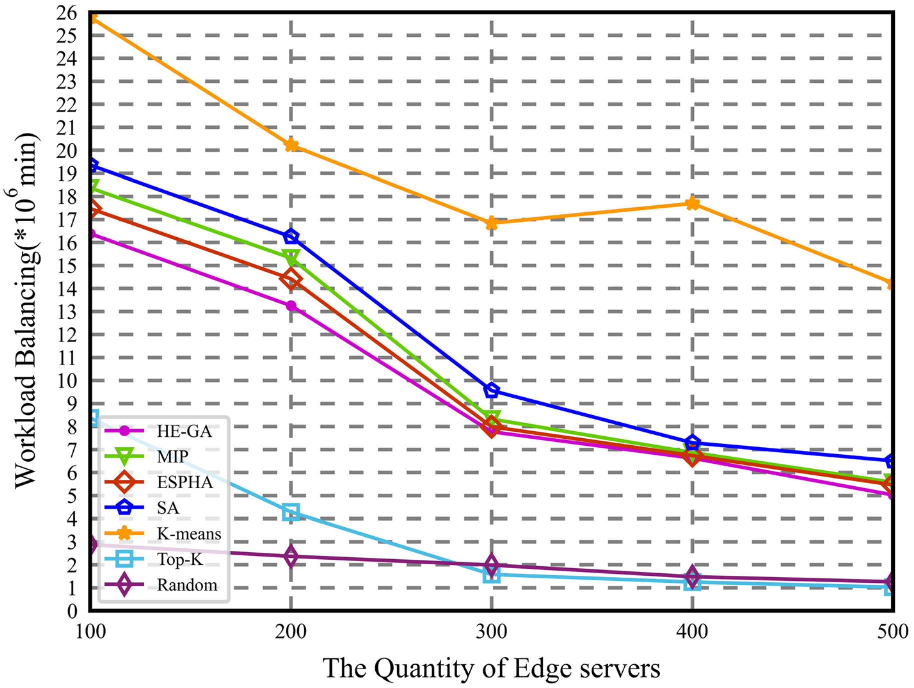

Fig. 6 Workload balancing v.s. the quantity of edge servers of fixed mobile communication base stations

data set to obtain the optimal deployment number of edge servers and consequently its value is 300. At this time, the deployment rate ER is equal to 0.1 and the benefit rate ρ is equal to 0.36794.

Tables 3, 4, 5, and 6 and Figs. 3 and 4 show the performance curve of edge server deployment under different edge server deployment methods during the number of mobile communication base stations increases from 300 to 3000 in the step of 300 when the deployment rate $ER = 0.1$. The abscissa in the figure indicates the quantity of mobile communication base stations, and the ordinate indicates the load balancing of edge server and communication delay. On the whole, in consideration of the optimal benefits, the HE-GA method raised in this paper has the best overall performance. Specifically speaking,

- In terms of communication delay, the HE-GA is the best. When the number of base stations is 3000, the communication delay of HE-GA is respectively 3.12%, 4%, 34.85%, 4.74%, 52.16%, and 58.04%, which is lower than those of ESPHA, MIP, SA, K-means, Random, and Top-K. The result indicates that under the constraint of optimizing the benefits, HE-GA achieves better communication delay compared to the baselines. Both Random and Top-K have high communication delay with significant fluctuations, and their overall performance is similar. However, when the number of base stations is 3000, the communication delay of Top-K more than that of Random by 3.87%. It is because that Shanghai, as an international metropolis, has a high population density and uneven distribution. Using the Top-K algorithm will give priority to base stations with a higher load, but ignores the load balance, which leads to higher communication delay.
- In terms of the standard deviation of edge server load, Random and Top-k perform best. When the number of base stations is 3000, the load standard deviation of HE-GA is respectively higher than those of Random and Top-K by 390.95%, 492.32%. At this point, the load standard deviation of Top-K is 25.93%, which is lower than that of Random. Because the two methods give priority to process base stations with a higher load, especially Top-K, so that the load gap of edge server is not large. In this case, the load standard deviation of HE-GA is respectively lower than those of ESPHA, MIP, SA, and K-means by 2.8%, 7.09%, 23.07%, and 116.75%. This result illustrates that under the constraints of optimal benefit, HE-GA achieves a better load standard deviation of edge server compared to the mainstream approach. Although Random and Top-K perform better in terms of load standard deviation, they do not consider optimal benefit and has poor communication delay performance. Therefore, their overall performance is average. K-means clusters the base stations with similar characteristics so that the base stations with similar load are more concentrated, that is to say, the base stations with the high load and low load are all more concentrated. However, the population density of Shanghai is too large and uneven, so the clustering effect is not ideal. Therefore, the K-means performs poorly in the comparative experiment of the load standard deviation of the edge server.
- Overall, HE-GA achieves better performance under the constraint of optimal benefit. The baselines, such as ESPHA, MIP, SA, K-means, Random, and Top-K, do not consider the constraint of optimal benefit, thus failing to effectively balance the conflicting interests of network operators and end-users. HE-GA considers optimal benefit, communication delay between edge servers and mobile communication base stations, and balances loads of edge servers simultaneously, thereby obtaining the overall optimal deployment solution.

Analysis of the results of different edge server numbers under the same quantity of base stations

To consider impact of different deployment rates on the load balance and communication delay of edge server, we make an experimental analysis without the optimal benefit quantity limit. Tables 7 and 8 and Figs. 5 and 6 show the performance curve of edge server deployment under different edge server deployment methods while the quantity of mobile communication base stations is determined at 3000 and the quantity of edge servers increases from 100 to 500 in footsteps of 100. The abscissa indicates the quantity of edge servers, and the ordinate indicates the load balancing and communication delay of edge servers. On the whole, without the constraint of optimal benefits, the HE-GA method still achieves the best communication delay under the acceptable standard deviation of load, and the entire performance is the optative. Specifically speaking,

- In the light of communication delay, the performance of the HE-GA is still the best on the whole. When the number of edge servers is 500, HE-GA achieves the reductions of communication latency by 11.75%, 19.84%, 42.56%, 28.74%, 64.64%, and 62.73% compared to ESPHA, MIP, SA, K-means, Random, and Top-K, respectively. This result indicates that even without the constraint of optimal benefits, HE-GA still achieves better performance in terms of communication latency compared to the baselines. Among

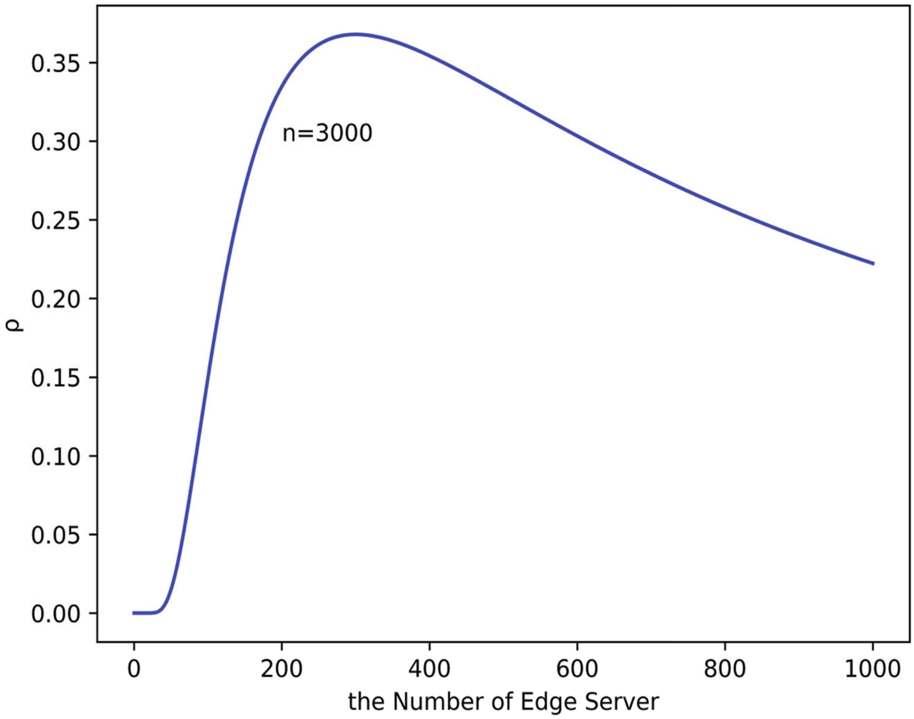

Fig. 7 Optimal benefit rate curve of edge servers

all the methods, the communication delays of Random and Top-k decrease the fastest by the increasing quantity of edge servers deployed. Because newly added edge servers greatly improve the imbalance of edge server deployment caused by the above two methods, thus greatly reducing the communication delay, but its performance is still lower than the HE-GA method. It is worth mentioning that before the quantity of edge servers reaches 300, the communication delay of edge servers for all deployment methods decreases significantly, and the decrease is higher than that of the process while the quantity of edge servers increases from 300 to 500.

- In the light of loads' standard deviation of edge servers, Random and Top-K still perform the best. When the number of edge servers is 500, the load standard deviation of HE-GA more than those of Random and Top-K by 398.87%, 490.04%, respectively. At this time, the load standard deviation of HE-GA is lower than ESPHA, MIP, SA, and K-means by 8.72%, 10.92%, 29.54%, and 183.53%, respectively. This result indicates that without the constraint of optimal benefits, HE-GA also achieves better load standard deviation of edge server compared to mainstream methods. It is also worth that in process of raising the quantity of edge servers from 100 to 300, except for Random, the loads' standard deviation of other deployment methods decreased significantly, and the decrease is higher than that of the process of raising the quantity of edge servers from 300 to 500.

On the whole, by the increasing quantity of edge server deployment, computing power and storage resources provided by the mobile service computing network also increase, so that each base station has the opportunity to allocate more computing power and storage resources to handle the service task requirements of its mobile intelligent terminal. The communication delay and loads' mean square error of the edge server will decrease rapidly with the raising of the quantity of edge servers, and the deployment effect of the edge server is better before reaching the optimal number of effective deployments. In general, the HE-GA has the best performance in most cases.

Analysis on deployment rate

Due to the lack of equipment hardware cost, communication loss, and other additional cost data of edge system overhead data in the Shanghai Telecom mobile communication base station data set, we set the additional cost as about 90% of the overall cost according to empirical data. Thus, we obtain the optimal benefit curve of edge servers when the quantity of base stations $n = 3000$, as Fig. 7 illustrates.

In Fig. 7, the ordinate represents the optimal benefit rate of edge servers, and abscissa is the quantity of edge servers deployed. Before the optimal benefit quantity curve reaches the highest point, by the quantity of deployed edge servers increasing, its deployment benefits are on the rise, and beyond the peak, if edge servers continue to be deployed, their benefits will gradually decline. Without considering the optimal benefit quantity curve,

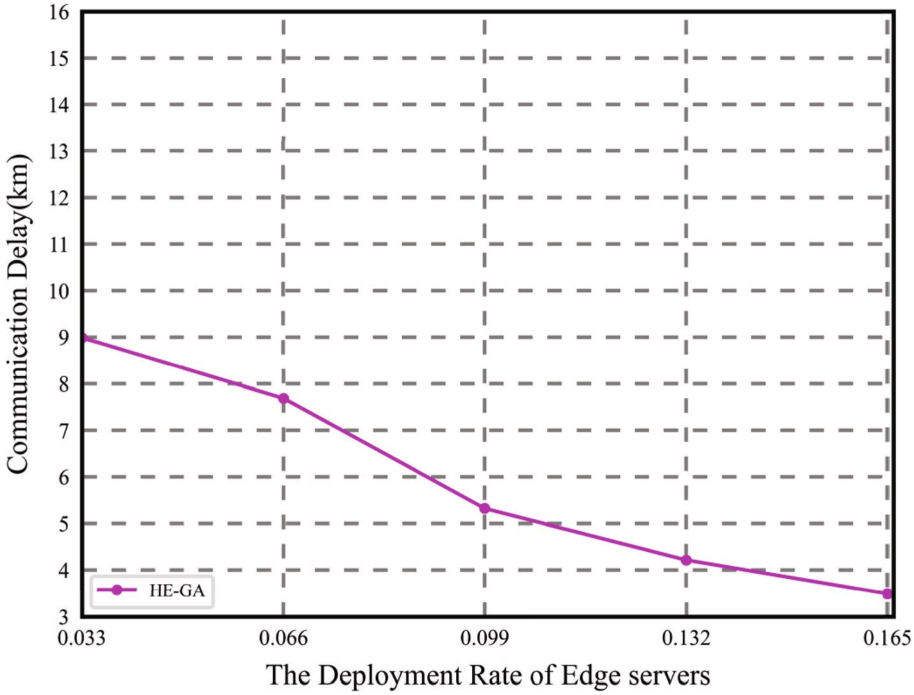

Fig. 8 Communication delay v.s. the edge server deployment rate

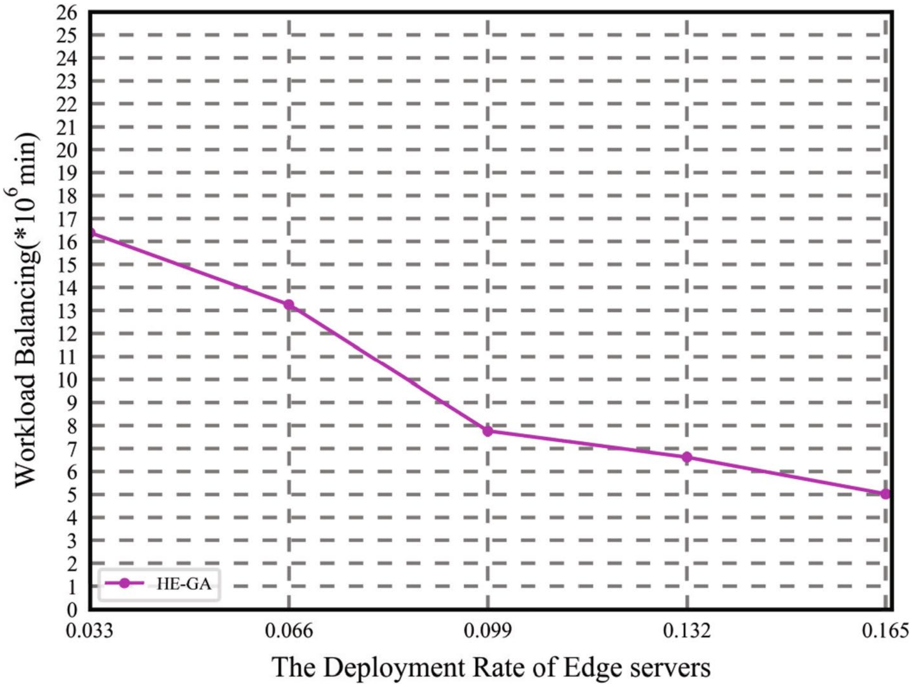

Fig. 9 Workload balancing v.s. the edge server deployment rate

the Figs. 8 and 9 show the changes of the performance of HE-GA model when the edge server deployment rate ER is gradually increased from 0.033 to 0.167. It is obvious from the two figures by the raising of the deployment rate of edge servers, the communication delay between edge servers and mobile communication base stations decreases, and the load difference between edge servers decreases too. It is worth noting that in process of increasing the deployment rate of edge server from 0.033 to 0.1, the communication delay and loads' standard deviation of HE-GA method have decreased significantly, and the decline is higher than that of edge server from 0.1 to 0.167. The reason is that before reaching the optimal benefit deployment rate, with the increase of the deployment rate of the edge server, the edge server will more effectively solve the problem of intercourse delay with the base station, and the service task requests will be more effectively distributed. After exceeding the optimal benefit deployment rate, even if we continue to increase investment, deploy more edge servers and improve the edge server deployment rate, the effect of improving service quality will gradually decrease, showing a state of diminishing benefit margin. This phenomenon fully shows the effectiveness of the optimal benefit quantity calculation method which we utilized.

Discussion

To simplify the calculations and focus on the load distribution of mobile communication base stations, this paper only considers the deployment problem in the scenario where edge servers are homogeneous. The investigation does not take into account how to obtain a more optimal deployment when edge servers have different computation and storage resources. In realistic scenarios, the edge servers deployed by operators may be homogeneous, the proposed method in this paper has certain practical significance. Meanwhile, the edge servers deployed by operators may also be heterogeneous. Therefore, when the computing and storage resources of edge servers are inconsistent, how to achieve the optimal deployment for edge servers is an interesting problem. In addition, the load required for mobile smart terminals to request services also has a significant impact on the deployment of edge servers. Given the joint request load of mobile smart terminals and mobile communication base stations, how to seek the optimal deployment for edge servers is also a problem worth to explore.

Conclusion and prospect

By the speedy advancement growth of 5G technology, edge server deployment has become a very important issue. An effective edge server deployment method can significantly reduce communication latency and energy consumption by integrating computing power and storage resources, thus it promotes the rapid development of mobile edge computing. In a typical mobile edge computing network environment, concentrating on the advancement of edge servers in case of dual-objective optimization of intercourse delay and load balancing under the optimal benefit quantity, we raise an edge server deployment method called HE-GA to achieve high-quality distributed deployment of edge servers. Compared to the mainstream edge server deployment methods, the HE-GA method applies the theory of optimal communication strategies in the communication field to the optimal benefit deployment strategy of edge servers, and achieves clear results in experiments. The experimental results show that the HE-GA validly counterpoises the edge server load and maximum lessens the communication delay between edge servers and base stations under the limitation of the optimal number of deployments. HE-GA method can be widely applied in the deployment of edge servers in smart city construction. In future work, we will apply deep learning algorithms, such as reinforcement learning, on the service deployment, offloading, and migration.

Acknowledgements
The authors would like to thank all the reviewers for their helpful comments.

Authors' contributions
Hongfan Ye: Conceptualization, Methodology, Software, Data curation, Validation, Writing-Original draft. Buqing Cao: Resources, Formal analysis, Funding acquisition, Supervision, Writing-Reviewing and Editing. Jianxun Liu: Investigation, Writing-Reviewing and Editing. Pei Li: Writing-Reviewing and Editing. Bing Tang: Writing-Reviewing and Editing. Zhenlian Peng: Writing-Reviewing and Editing.

Funding
Our work is supported by the National Natural Science Foundation of China (No. 61873316), the National Key R &D Program of China (No. 2018YFB1402800), the Hunan Provincial Natural Science Foundation of China (No. 2021JJ30274), the Postgraduate Scientific Research Innovation Project of Hunan Province (No. CX20221045), the Hunan Provincial Natural Science Foundation of China (No. 2021JJ30283), the Hunan Provincial Natural Science Foundation of China (No. 2021JJ30278), the Scientific Research Fund of Hunan Provincial Education Department (No. 22B0497), and the Educational Commission of Hunan Province of China (No.20A175). Buqing Cao is the corresponding author of this paper.

Availability of data and materials
The datasets generated and material during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The research work in this paper does not involve any human or animal studies.

Competing interests
The authors declare no competing interests.

Received: 31 January 2023 Accepted: 27 September 2023
Published online: 18 October 2023

References

1. Fernando N, Loke SW, Rahayu W (2013) Mobile cloud computing: A survey. Futur Gener Comput Syst 29(1):84–106
2. Andrews JG, Buzzi S, Choi W, Hanly SV, Lozano A, Soong AC, Zhang JC (2014) What will 5g be? IEEE J Sel Areas Commun 32(6):1065–1082
3. Neirotti P, De Marco A, Cagliano AC, Mangano G, Scorrano F (2014) Current trends in smart city initiatives: Some stylised facts. Cities 38:25–36
4. Deng S, Huang L, Wu H, Tan W, Taheri J, Zomaya AY, Wu Z (2016) Toward mobile service computing: Opportunities and challenges. IEEE Cloud Comput 3(4):32–41
5. Dinh HT, Lee C, Niyato D, Wang P (2013) A survey of mobile cloud computing: architecture, applications, and approaches. Wirel Commun Mob Comput 13(18):1587–1611
6. Shi W, Cao J, Zhang Q, Li Y, Xu L (2016) Edge computing: Vision and challenges. IEEE Internet Things J 3(5):637–646
7. Shi W, Zhang X, Wang Y, Zhang Q (2019) Edge computing: state-of-the-art and future directions. J Comput Res Dev 56(1):69–89
8. Zhao Z, Liu F, Cai Z, Xiao N (2018) Edge computing: platforms, applications and challenges. J Comput Res Dev 55(2):327–337
9. Cong R, Zhao Z, Min G, Feng C, Jiang Y (2021) Edgego: A mobile resource-sharing framework for 6g edge computing in massive iot systems. IEEE Internet Things J 9(16):14521–14529
10. Mao Y, You C, Zhang J, Huang K, Letaief KB (2017) A survey on mobile edge computing: The communication perspective. IEEE Commun Surv Tutorials 19(4):2322–2358
11. Zhang P, Jin H, Dong H, Song W, Bouguettaya A (2020) Privacy-preserving qos forecasting in mobile edge environments. IEEE Trans Serv Comput 15(2):1103–1117
12. Wang X, Li J, Ning Z, Song Q, Guo L, Guo S, Obaidat MS (2023) Wireless powered mobile edge computing networks: a survey. ACM Comput Surv 55(13):1–37
13. Hu YC, Patel M, Sabella D, Sprecher N, Young V (2015) Mobile edge computing—a key technology towards 5g. ETSI White Pap 11(11):1–16
14. Zhang P, Zhang Y, Dong H, Jin H (2020) Mobility and dependence-aware qos monitoring in mobile edge computing. IEEE Trans Cloud Comput 9(3):1143–1157
15. Qi Q, Wang J, Ma Z, Sun H, Cao Y, Zhang L, Liao J (2019) Knowledge-driven service offloading decision for vehicular edge computing: A deep reinforcement learning approach. IEEE Trans Veh Technol 68(5):4192–4203
16. Shi C, Habak K, Pandurangan P, Ammar M, Naik M, Zegura E (2014) Cosmos: computation offloading as a service for mobile devices. In: Proceedings of the 15th ACM international symposium on Mobile ad hoc networking and computing. ACM, New York, pp 287–296
17. Xie R, Lian X, Jia Q, Huang T, Liu Y et al (2018) Survey on computation offloading in mobile edge computing. J Commun 39(11):138–155
18. Chen M, Li W, Fortino G, Hao Y, Hu L, Humar I (2019) A dynamic service migration mechanism in edge cognitive computing. ACM Trans Internet Technol (TOIT) 19(2):1–15
19. Wang S, Urgaonkar R, Zafer M, He T, Chan K, Leung KK (2015) Dynamic service migration in mobile edge-clouds. In: 2015 IFIP Networking Conference (IFIP Networking). IEEE, Toulouse, pp 1–9
20. Zhang K, Gui X, Ren D, Li J, Wu J, Ren D (2019) Review of computing migration and content caching in mobile edge networks. J Softw 8:2491–2516
21. Wang S, Xu J, Zhang N, Liu Y (2018) A survey on service migration in mobile edge computing. IEEE Access 6:23511–23528
22. Sjödin D, Parida V, Kohtamäki M, Wincent J (2020) An agile co-creation process for digital servitization: A micro-service innovation approach. J Bus Res 112:478–491
23. Li Q, Wang S, Zhou A, Ma X, Yang F, Liu AX (2020) Qos driven task offloading with statistical guarantee in mobile edge computing. IEEE Trans Mob Comput 21(1):278–290
24. Kuang L, Gong T, OuYang S, Gao H, Deng S (2020) Offloading decision methods for multiple users with structured tasks in edge computing for smart cities. Futur Gener Comput Syst 105:717–729
25. Ma X, Zhou A, Zhang S, Wang S (2020) Cooperative service caching and workload scheduling in mobile edge computing. In: IEEE INFOCOM 2020-IEEE Conference on Computer Communications. IEEE, pp 2076–2085
26. Xu J, Ma X, Zhou A, Duan Q, Wang S (2020) Path selection for seamless service migration in vehicular edge computing. IEEE Internet Things J 7(9):9040–9049
27. Huang L, Feng X, Feng A, Huang Y, Qian LP (2018) Distributed deep learning-based offloading for mobile edge computing networks. Mob Netw Appl 27:1123–1130
28. Huang L, Bi S, Zhang YJA (2019) Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks. IEEE Trans Mob Comput 19(11):2581–2593
29. Wang H, Li Y, Zhou A, Guo Y, Wang S (2020) Service migration in mobile edge computing: A deep reinforcement learning approach. Int J Commun Syst 36(1):e4413
30. La QD, Ngo MV, Dinh TQ, Quek TQ, Shin H (2019) Enabling intelligence in fog computing to achieve energy and latency reduction. Digit Commun Netw 5(1):3–9
31. Li Y, Wang S (2018) An energy-aware edge server placement algorithm in mobile edge computing. In: 2018 IEEE International Conference on Edge Computing (EDGE). IEEE, San Francisco, pp 66–73
32. Wang S, Zhao Y, Xu J, Yuan J, Hsu CH (2019) Edge server placement in mobile edge computing. J Parallel Distrib Comput 127:160–168
33. Chen Y, Lin Y, Zheng Z, Yu P, Shen J, Guo M (2021) Preference-aware edge server placement in the internet of things. IEEE Internet Things J 9(2):1289–1299
34. Davis L (1991) Handbook of genetic algorithms. IEEE Trans Cloud Comput (1991):1–101
35. Wang P, Xu J, Zhou M, Albeshri A (2023) Budget-constrained optimal deployment of redundant services in edge computing environment. IEEE Internet Things J 10(11):9453–9464
36. Chang L, Deng X, Pan J, Zhang Y (2021) Edge server placement for vehicular ad hoc networks in metropolitans. IEEE Internet Things J 9(2):1575–1590
37. Cruz P, Achir N, Viana AC (2022) On the edge of the deployment: A survey on multi-access edge computing. ACM Comput Surv 55(5):1–34
38. Fan W, Chen Z, Hao Z, Su Y, Wu F, Tang B, Liu Y (2022) Dnn deployment, task offloading, and resource allocation for joint task inference in iiot. IEEE Trans Industr Inf 19(2):1634–1646
39. Kasi SK, Kasi MK, Ali K, Raza M, Afzal H, Lasebae A, Naeem B, Ul Islam S, Rodrigues JJ (2020) Heuristic edge server placement in industrial internet of things and cellular networks. IEEE Internet Things J 8(13):10308–10317
40. Zhang X, Li Z, Lai C, Zhang J (2021) Joint edge server placement and service placement in mobile-edge computing. IEEE Internet Things J 9(13):11261–11274
41. Asghari A, Sohrabi MK (2022) Multiobjective edge server placement in mobile-edge computing using a combination of multiagent deep q-network and coral reefs optimization. IEEE Internet Things J 9(18):17503–17512
42. Yin H, Zhang X, Liu HH, Luo Y, Tian C, Zhao S, Li F (2016) Edge provisioning with flexible server placement. IEEE Trans Parallel Distrib Syst 28(4):1031–1045
43. Cui G, He Q, Chen F, Jin H, Yang Y (2020) Trading off between user coverage and network robustness for edge server placement. IEEE Trans Cloud Comput 10(3):2178–2189
44. Guo Feiyan TB (2021) Mobile edge server placement method based on user delay perception. Comput Sci 48(1):103–110
45. Ji X, He Y, Wang J, Dong W, Wu X, Liu Y (2014) Walking down the stairs: Efficient collision resolution for wireless sensor networks. In: IEEE INFOCOM 2014-IEEE Conference on Computer Communications. IEEE, pp 961–969
46. Charikar M, Guha S, Tardos É, Shmoys DB (2002) A constant-factor approximation algorithm for the k-median problem. J Comput Syst Sci 65(1):129–149

Publisher's Note