This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3328595

Cost and Delay-Aware Service Replication for Scalable Mobile Edge Computing

Shimaa A. Mohamed, Sameh Sorour, Senior Member, IEEE, Sara A. Elsayed, Member, IEEE, and Hossam S. Hassanein, Fellow, IEEE

***Abstract*—Mobile Edge Computing (MEC) has emanated as a propitious computing paradigm that can foster delay-sensitive and/or data-intensive applications. However, it can be challenging to maintain a scalable MEC service when computational resources are overloaded. In this paper, we propose the Service Replication between Multiple Service Providers (SRMSP) scheme. SRMSP is the first scheme that fosters service scalability in a cost-efficient manner, while considering the stringent QoS requirements of real-time applications involving groups of users. SRMSP enables service replication between multiple service providers to minimize the average response delay and the operational cost incurred by service providers, while satisfying the delay requirements of all user groups. We formulate the resource allocation problem as an Integer Linear Program (ILP) and derive an analytical solution using the Karush-Kuhn-Tucker (KKT) conditions and Lagrangian analysis. In addition, we propose the SRMSP-Distributed Allocation (SRMSP-DA) scheme to provide a time-efficient solution in distributed scenarios. In SRMSP-DA, we use a game-theoretic strategy that formulates the resource allocation problem as a potential game. Extensive simulations show that SRMSP renders a 50% operational cost reduction compared to a baseline scheme that does not consider the operational cost. In addition, SRMSP-DA exhibits a relatively marginal difference of up to 20% and 4% in terms of the total operational cost and average response delay, respectively, compared to the optimal solution provided by SRMSP.**

***Index Terms*—Mobile Edge Computing, Service Providers, Resource Allocation, Service Replication, Lagrangian Analysis**

I. Introduction

The unprecedented proliferation of IoT applications that hinge on synchronized experiences and stringent Quality of Service (QoS) requirements for groups of users is expected to impose tremendous demands on computing resources [1]. Online gaming and live sports events and concerts, which require real-time synchronization, find themselves at the forefront of these challenging demands. With the high-density workloads of such applications, it is imperative to have efficient computing solutions that not only maintain strict QoS requirements and real-time synchronization but also ensure high scalability [2].

Mobile Edge Computing (MEC) has emerged as a promising computing paradigm that can bring the computing service closer to end-users, thus drastically curtailing the delay and improving the QoS [3]. However, despite its potential, MEC relies on edge servers with limited computational resources compared to centralized cloud servers [4], [5]. These limitations, particularly in situations characterized by high-density workloads and rigorous Quality of Service (QoS) requirements, can lead to server overloads and the risk of service interruptions for some users [6]. In this context, Service Providers (SPs) strive to efficiently manage their MEC servers to maintain a high QoS and improve scalability to serve as many clients as possible, while simultaneously reducing operational costs. Note that operational costs encompass the expenses associated with providing computing resources, services, and support at the network's edge [7]. Simply increasing the number of MEC servers or computational resources may improve scalability, increase the number of clients served, and enhance the QoS. However, this approach can also escalate operational costs, potentially diminishing the SPs' overall profitability [8]. Thus, it is crucial to make resource allocation decisions that strike a balance between ensuring scalability, maintaining high QoS, and minimizing operational costs.

Existing resource allocation schemes mostly focus on achieving certain QoS objectives, such as efficient resource utilization [9], reduced latency [10]–[12], and increased energy savings [13]–[15]. However, such schemes fail to address the issue of scalability in MEC systems, thus increasing the risk of service interruptions and intermittent availability. To ensure service continuity under excessive computation loads, some schemes rely on service migration [16], while others resort to service replication [17]. However, these schemes overlook the associated operational costs and fail to accommodate the stringent QoS requirements imposed by applications that target groups of users. Thus, existing resource allocation schemes fail to address the challenging issue of balancing between ensuring scalability to effectively manage overloaded servers, reducing operational costs, and satisfying the strict QoS requirements for real-time services requiring user synchronization.

To address the aforementioned challenge, we propose the Service Replication between Multiple Service Providers (SRMSP) scheme. SRMSP considers multiple sets of users forming service groups based on each group's participation in the same synchronized activity, such as online gaming, multimedia conferencing, and augmented reality. SRMSP endeavors to improve the scalability of such real-time services by replicating them on-demand at underloaded edge servers of foreign SPs to serve all users concurrently when home SPs fail to sustain the load from all user groups. Note that

Shimaa A. Mohamed is with the School of Computing, Queen's University, Kingston, ON, Canada and with Network and Distributed Systems Department, Informatics Research Institute (IRI), the City of Scientific Research and Technological Applications (SRTA-City), Alexandria, Egypt, Email: s.abdelmeguid@queensu.ca

Sameh Sorour, Sara A. Elsayed and Hossam S. Hassanein are with the School of Computing, Queen's University, Kingston, ON, Canada, Email: sameh.sorour@queensu.ca, selsayed@cs.queensu.ca, hossam@cs.queensu.ca

foreign SPs are SPs that the user/requester is not subscribed to, whereas a home SP is the one to which the user is subscribed (i.e., has a monthly or yearly contract with). Based on the SPs' predefined cost information, we dynamically manage resource allocation for groups of users requesting to perform inter-related computations between different service providers; home service providers, and foreign service providers. Such resource allocation decisions are centrally managed.

SRMSP is a centralized scheme that involves an orchestrator that is responsible for making all the necessary decisions. We formulate the resource allocation problem as an Integer Linear Program (ILP) and provide an analytical solution using the Karush-Kuhn-Tucker (KKT) conditions and Lagrangian analysis. In order to deal with distributed scenarios, where there is no orchestrator, we also propose the SRMSP-Distributed Allocation (SRMSP-DA) scheme.

The main contributions of this paper are as follows:

- We propose a novel scheme, SRMSP, that accounts for and balances between service scalability and minimizing the operational cost of SPs, as well as sustaining a certain QoS. SRMSP is the first scheme that performs service replication at underloaded foreign SPs in a way that minimizes the average response delay and the operational cost incurred by overloaded home SPs, while satisfying the strict QoS requirements of user groups that run the same synchronized application.
- We solve the resource allocation problem in a distributed and time-efficient manner by using a heuristic scheme that employs a game-theoretic approach. The proposed SRMSP-DA scheme formulates the Service Replication Game (SRG) to reach a collectively satisfactory resource allocation solution. SRMSP-DA scales with the size of the SRG, including the number of home service providers and their local users, as well as the number of available foreign service providers.

We evaluate the performance of SRMSP and SRMSP-DA in comparison to a baseline scheme that prioritizes delay without considering operational costs [18], as well as a greedy approach that offloads users' tasks to the nearest foreign service provider [7]. Extensive simulations demonstrate that SRMSP significantly outperforms both the former and the latter, achieving improvements of up to 50% and 54% in terms of the total operational cost, respectively, while maintaining a small gap of up to 4% and 4%, respectively, in terms of the average response delay. Moreover, SRMSP-DA closely approaches the optimal solution achieved by SRMSP, with discrepancies of up to 20% in the total operational cost and up to 4% in the average response delay.

The rest of the paper is organized as follows; Section II presents an overview of the related work. Section III introduces our proposed SRMSP scheme. Section IV presents the SRMSP-DA scheme. Section V illustrates the performance evaluation and simulation results. Section VI highlights our conclusions and future research directions.

II. Related Work

Most of the existing resource allocation schemes focus on optimizing certain QoS metrics, such as efficient resource utilization, latency, and energy consumption. In [9], Kumar et al. proposed a scheme that strives to optimize the resource utilization of edge servers deployed at the network's edge. They used a game-theoretic approach and formulated the Edge Resource Allocation (ERA) problem as a potential game to optimize resource utilization. In [10], Shabir et al. proposed a fully distributed task offloading framework between multiple vehicular destinations to improve the resource utilization efficiency of the vehicular edge and the overall delay, including the queuing delay and data transmission delay. The task offloading framework replicated the tasks at the neighboring vehicles and selected the most suitable vehicle based on the response time of all vehicles in the previous time slot. In [11], Zhu et al. studied the use of Intelligent Reflecting Surface (IRS) in MEC-served vehicular networks. They optimized the latency of task scheduling by enhancing the allocation of limited processors and IRS resources. They proposed a dynamic task scheduling algorithm that takes into account both communications and computations, considering vehicle mobility patterns, transmission conditions, and task sizes. In [12], Rodrigues et al. conducted an analysis of learning costs, focusing on the transmission and processing delays associated with both centralized and distributed learning methods. They proposed a hybrid solution that intelligently leverages the strengths of both approaches within a satellite network equipped with cloud servers. In this context, they introduced a hybrid approach that combines centralized and distributed learning techniques to train a deep Q network (DQN) model in a satellite network enhanced with MEC capabilities. The findings indicated that the optimal strategy involves capitalizing on the advantages of both centralized and distributed paradigms when necessary, while also remaining adaptable to scenario-specific factors affecting transmission and processing delays. In [13], Verma et al. proposed a novel energy-efficient grouping method for mobile IoT devices, taking into consideration the packet arrival delay and the packet loss rate. They formulated an optimization problem to minimize the energy consumption of mobile IoT devices. In [14], Guo et al. conducted a study on the utilization of unmanned aerial vehicles (UAVs) to enhance edge computing energy efficiency. Their approach involves making intelligent offloading decisions, efficiently allocating transmitted data in both uplink and downlink directions, and designing UAV trajectories. They formulated a joint optimization problem and proposed an intelligent approach that leverages block coordinate descent and successive convex approximation techniques. In [15], Guo et al. introduced a hybrid fiber wireless (FiWi) network concept designed to facilitate the coexistence of centralized cloud and Multi-access Edge Computing (MEC). They outlined an architectural framework that incorporates FiWi access networks for this purpose. Additionally, they delved into the challenge of collaborative computation offloading between cloud and MEC to minimize the total energy consumption while satisfying the maximum execution latency.

The aforementioned schemes focused on improving certain QoS metrics, such as delay and energy. However, they did not consider the operational cost of edge computing servers. In [8], Zhang et al. aimed to minimize the operational cost of

an edge computing system consisting of multiple collocated edge servers by formulating the resource allocation problem of delay-sensitive tasks as a bin packing problem. They proposed a two-stage task scheduling algorithm to solve the cost-based optimization problem. In [19], Ma et al. proposed a cloud-assisted mobile edge computing solution to balance the load between the cloud and MEC servers and minimize the total users' delay, internal cost (i.e., infrastructure), outsourcing cost, and cloud computation. In [20], Herrera et al. proposed a Next-gen IoT Optimization (NIoTO) framework to optimize the placement of micro-services and networking resources for optimizing the QoS considering the computing, networking, and application dimensions. They focused on response time and deployment cost (i.e., the CAPEX and the OPEX) because reducing the response time required a higher cost, thus, a tradeoff between them is challenging.

While they focus on delay, energy, and operational cost, most existing resource allocation schemes fail to account for scalability in MEC systems, which leads to high risks of service interruptions and intermittent availability [6]. In [21], Liu et al. proposed a joint optimization objective for the unavailability level of the edge servers and switches, the communication delay, and the resource wastage while allocating the same batch of IoT applications to multiple edge clouds. In [16], Zaki et al. proposed the Dynamic Load-based Proactive Migration (DLPM) scheme to deal with service interruptions. They proactively migrated the service to the optimal nearest edge node with the highest probability of having the lowest computational load. However, this approach could lead to excessive migration delay in case of erroneous load estimations. In addition, it assumed that all edge nodes were managed by the same SP. Rather than resorting to service migration, Mohamed et al. [17] improved service scalability by replicating services at distant edge servers when the local server(s) could not sustain the load from all user groups by assuming that all edge servers belong to a single SP. In addition, the operational cost was not considered, and the strict QoS requirements imposed by applications that involve a group of users were not sufficiently satisfied. In [18], the Delay-Oriented Service Replication (DOSR) scheme was proposed to enhance the concept of user grouping by introducing the time and time different constraints for each group to improve the scalability of real-time applications that could not tolerate queuing at MEC servers. However, this scheme did not consider the operational cost incurred by overloaded SPs to be served by underloaded SPs.

In contrast to existing schemes, we account for the scalability of MEC systems that have multiple SPs by replicating the service at underloaded foreign SPs, while minimizing the operational cost and maintaining a certain QoS. We replicate the service at underloaded foreign SPs that yield the minimum average response delay and the minimum operational cost incurred by overloaded home SPs. This replication is done is while satisfying the delay requirements of user groups running the same application.

III. Service Replication between Multiple Service Providers (SRMSP)

In this section, we provide a detailed discussion of the system model, the problem formulation, and the analytical solution of SRMSP.

A. System Model and Overview

In SRMSP, the MEC system consists of a set of edge servers and a set of users requesting a service/task that needs to be offloaded to an edge server for execution. Several edge service providers manage the edge servers. For simplicity, we assume that each service provider manages one edge server in the MEC system. As shown in Figure 1, the edge service providers are divided into two subsets; home and foreign edge service providers. Each home service provider provides the required service to its users without any additional cost unless overloaded. Such an additional cost stems from the fact that foreign service providers invoice the home service provider for serving unserved users. In SRMSP, the system is characterized by the following:

- Foreign service providers have underutilized computing resources.
- Home service providers have demands that exceed their capacities. Thus, unserved users of home SPs are allowed to access the foreign SPs' resources after the services of home SPs have been replicated to foreign SPs.
- It is known from the mobile network that mobile users can connect to other service providers for roaming purposes when they are far from the server/base station belonging to their contracted service provider [22]. We use the same concept for MEC purposes. Whenever a user cannot access the computing service via the edge server of its home SP, each user can wirelessly connect to any available nearby edge server belonging to a foreign SP. This is while considering the different levels of connection qualities due to fading conditions.

Let $S^h = \{s_1^h, s_2^h, \ldots, s_m^h\}$ be the set of home service providers in the system, $m = |S^h|$ denotes the number of overloaded home service providers, $S^f = \{s_1^f, s_2^f, \ldots, s_k^f\}$ be the set of foreign service providers in the system, $k = |S^f|$ denotes the number of underloaded foreign service providers in the system, and $S = s_j^h \bigcup S^f = \{s_1, s_2, \ldots, s_{k+1}\}$ be the set of possible service providers that the user can offload its tasks to. All edge servers belonging to home service providers and foreign service providers in the MEC system are connected via backhaul links, which are used for service replication when needed. In addition, the provider's computational resources are divisible among multiple users. Let C_j^h denote the available computing capacity (in GHz) of each home edge server $s_j^h \in S^h$, and C_j^f denote the available computing capacity (in GHz) of each foreign edge server $s_j^f \in S^f$ in the MEC system.

At the user side, let $G = \{G_1, G_2, \ldots, G_z\}$ be the set of groups of users that participate in the same data-intensive, real-time, and multimedia-oriented application/activity, where

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3328595

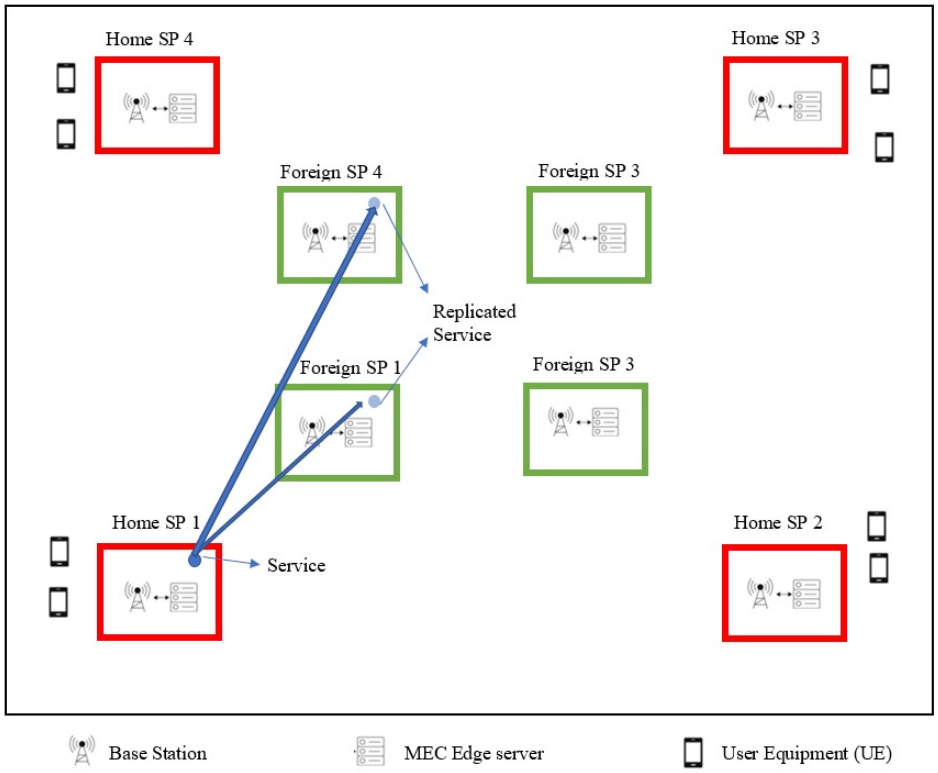

Fig. 1: System model of SRMSP

each group $G_g = \{u_1^g, u_2^g, \ldots, u_{y_g}^g\}$[1] represents the set of users sharing the same application/activity inside Group G_g. Let $U = \bigcup_{g=1}^{z} G_g = \{u_1, u_2, \ldots, u_n\}$ be the set of users subscribing to the system, and U^j be the set of n_j users subscribing to the home service provide s_j^h, where $U^j \subset U$. Being involved in the same application, members of each group are usually constrained with application-dependent QoS requirements. T_g^{max}, F_g^{min}, and d_g denote the maximum delay that all the members of Group G_g can tolerate to complete the task, the minimum CPU speed needed for the task requested by Group G_g to be executed within the specified deadline T_g^{max}, and the maximum difference between the delay experienced between any two users belonging to the same Group G_g (i.e., the maximum delay difference), respectively. Note that users must get the response of their offloaded computation within a minimal difference in time, or else the application fails. An example of this type of application is online gaming, in which players must see the results of each other's actions almost at the same time or within a bounded time difference. Otherwise, different players can end up seeing other users' actions with various delays, which may cause some players to make wrong decisions in their following actions/moves, thus totally ruining the game.

Each user sends its task to the MEC server of its home service provider or one of its accessible foreign service providers. We identify a user's task $\mathcal{T}_{ig}$ by the 3-tuples (c_i, b_i, G_g), as given by Eq. (1).

$$\mathcal{T}_{ig} = (c_i,\ b_i,\ G_g) \tag{1}$$

Note that $u_i^g \in G_g$ represents the user requesting the task, c_i denotes the workload intensity of the task (in Mega CPU cycles), b_i (in megabytes) describes the size of the information that User u_i^g needs to send to the edge server, and G_g denotes the ID of the group that User u_i^g belongs to. In our model, each user group is dedicated to a single service, and within the group, each user's task is associated with a specific service function deployed at the home service provider's edge server. The process of mapping conventional MEC service functions to a task involves identifying the precise sequence or chain of service functions that must be executed to fulfill a particular task or service request from an end-user or device [23], [24]. Each foreign service provider $s_j^f \in S^f$ has two-unit costs; τ_j^C, and τ_j^T, which are the unit cost of the computation rate per one Mcycle, and the unit cost of the transmission rate per one MByte, respectively. In addition, τ_g^R is a group-dependent parameter, denoting the unit cost of replicating the service of Group G_g on foreign service providers.

B. Problem Formulation

In SRMSP, our objective is to jointly minimize the average response delay and the fees paid by home service providers when users are compelled to send their tasks to foreign service providers. This should be achieved while ensuring that all Quality of Service (QoS) requirements of the group's users are met. The term response delay refers to the total delay experi-

[1]This model captures a special case scenario where one or multiple users are solely involved in applications. In this case, each of these users forms a separate group of size 1 (i.e., $y_g = 1$)

enced starting from the time a request is issued until a response is received. The response delay includes the transmission delay of the task to the edge server, the computation delay at the server, the replication delay of the application on a foreign SP's edge server (if needed), and the roaming time when a user accesses the required service(s) from a foreign SP. Similar to other works, we assume that the size of the result is too small that its corresponding transmission delay is negligible [25]. Note that the total cost includes the computation cost, transmission cost, and replication cost.

As the case in typical MEC settings, users offload their tasks and access the service at their assigned edge server through the direct wireless channel between them. The rate at which the instructions and data of User u_i^g's computational task $\mathcal{T}_{ig}$ is sent to Service Provider $s_j \in S$ is denoted R_{ij}. Consequently, the transmission time of Task $\mathcal{T}_{ig}$ of User u_i^g to Service Provider s_j is given by Eq. (2).

$$T_{ij}^T = \frac{b_i}{R_{ij}} \tag{2}$$

In addition, the computation time of task $\mathcal{T}_{ig}$ at Service Provider s_j is given by Eq. (3), where $f_{ig} \geq F_g^{min}$ is the assigned CPU frequency assigned to Task $\mathcal{T}_{ig}$, whose requester belongs to Group $G_g \subset G$ when it is offloaded to Service Provider s_j.

$$T_{ij}^C = \frac{c_i}{f_{ig}} \tag{3}$$

In addition, if $\mathcal{T}_{ig}$ is offloaded to foreign Service Provider $s_j^f \in S^f$, an additional time T_{ij}^R will be required to replicate and deploy this service on that foreign service provider's edge server. Note that the replication time depends on the required service to Group G_g that User $u_i^g \in U$ belongs to. Finally, the roaming time T_{ijg}^{RM} is the time needed to connect User u_i^g with other foreign service providers affiliated with the users' home service provider. Thus, the response time of $\mathcal{T}_{ig}$ at Service Provider s_j is denoted T_{ijg}^{resp} and is given by Eq. (4).

$$T_{ijg}^{resp} = T_{ij}^T + T_{ij}^C + T_{ij}^R + T_{ijg}^{RM} \tag{4}$$

We formulate the multi-objective task offloading problem as a 0-1 Integer Linear Program (0-1 ILP), where the decision variable x_{ij} is set to 1 if User $u_i^g \in U$ offloads its task to Service Provider $s_j \in S$, and 0 otherwise, and the decision variable b_{gj} is set to 1 if at least one user in Group G_g is granted a service from the foreign Service Provider s_j^f, and 0 otherwise. The problem formulation is given by Eq. (5).

$$\min_{x_{ij},\, b_{gj}} \{T^{resp}, C_{SP}\} \tag{5a}$$

$$S.T.$$

$$\sum_{s_j \in S} x_{ij} = 1, \qquad \forall u_i \in U, \tag{5b}$$

$$\sum_{u_i^g \in U^j} x_{ij} f_{ig} \leq C_j^h, \qquad \forall s_j^h \in \mathcal{S}^h, \tag{5c}$$

$$\sum_{u_i^g \in U} x_{ij} f_{ig} \leq C_j^f, \qquad \forall s_j^f \in S^f \tag{5d}$$

$$\sum_{s_j \in S} x_{ij} T_{ijg}^{resp} \leq T_g^{max}, \qquad \forall G_g \subset G, \forall u_i^g \in U \tag{5e}$$

$$\frac{\sum_{u_i^g \in G_g} x_{ij}}{y_g} - b_{gj} \leq 0, \quad \forall G_g \subset G, \forall s_j^f \in S^f \tag{5f}$$

$$b_{gj} - \sum_{u_i^g \in G_g} x_{ij} \leq 0, \qquad \forall G_g \subset G, \forall s_j^f \in S^f \tag{5g}$$

$$\sum_{s_j \in S} x_{ij} T_{ijg}^{resp} - \sum_{s_j \in S} x_{lj} T_{ljg}^{resp} \leq d_i,$$
$$\forall G_g \subset G,\ \forall u_i^g, u_l^g \in G_g,\ l \neq i \tag{5h}$$

$$x_{ij} \in \{0,1\},\ \forall s_j \in S,\ \forall u_i^g \in U \tag{5i}$$

$$b_{gj} \in \{0,1\}, \forall s_j^f \in S^f, \forall G_g \subset G \tag{5j}$$

Note that T^{resp} represents the average response delay and C_{SP} represents the total cost that all home service providers incur by all foreign service providers for serving all unserved users of the set U concurrently, as given by Eq. (6) and Eq. (7), respectively, where $C_j^S = (c_i \times \tau_j^C) + (b_i \times \tau_j^T)$ is the service cost (computation plus transmission).

$$T^{resp} = \frac{1}{n} \sum_{g=1}^{z} \sum_{u_i^g \in U^j} \sum_{s_j \in S} x_{ij} T_{ijg}^{resp} \tag{6}$$

$$C_{SP} = \sum_{u_i \in U} \sum_{s_j \in S^f} C_j^S * x_{ij} + \sum_{s_j \in S^f} \sum_{g=1}^{z} (b_{gj} * \tau_g^R) \tag{7}$$

Constraint (5b) ensures that each task is offloaded to only one service provider. Constraint (5c) ensures that the sum of assigned CPU frequencies to the users offloading their tasks to their home service provider does not exceed its total available CPU frequency. Constraint (5d) ensures that the sum of assigned CPU frequencies to all users offloading their tasks to the foreign service provider does not exceed its total available CPU frequency. Constraint (5e) ensures that the total response time of a user's task belonging to any Group $G_g \subset G$ does not exceed the service's maximum delay requirement. Constraints (5f) and (5g) ensure that the group's service is deployed on the edge server of the foreign service provider (i.e., ensuring that $b_{gj} = 1$ if there is at least one User u_i^g of Group G_g that accesses its service at foreign Service Provider s_j^f, and $x_{ij} = 1$). Constraint (5h) ensures that the delay difference between two users belonging to the same Group G_g does not exceed the predefined difference delay requirement. Constraint (5i) and (5j) impose a binary decision value for the variables x_{ij}, and b_{gj} of all users and services, respectively.

It is clear that the optimization problem given by Eq. (5) is an integer linear multi-objective problem. Note that the solution of the constrained multi-objective problem is characterized by a Pareto front formed by a set of possible solutions. Each solution represents a different trade-off between the average response delay of all users and the cost incurred by the

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3328595

Algorithm 1: Greedy Rounding

Input : x^*_{ij}, $\forall s_j \in S,\ \forall u^g_i \in U$, and b^*_{gj},$\forall s^f_j \in S^f,\ \forall G_g \subset G$
Output: The integer linear decision variables $x_{ij},\ \forall u^g_i \in U,\ \forall s_j \in S$, and b_{gj},$\forall s^f_j \in S^f,\ \forall G_g \subset G$
Steps:
Sort the set of decision variables x^*_{ij}, $\forall s_j \in S,\ \forall u^g_i \in U$ in a descending order
Rounding $x^*_{ij} \in \{0, 1\}$
if $round(x^*_{ij}) = 1$ **then**
  if *constraints (5b)-(5h) are achieved* **then**
    $x_{ij} \leftarrow 1$
  else
    $x_{ij} \leftarrow 0$
  end
else
  $x_{ij} \leftarrow 0$
end
/* Based on above decision variables $X = [x_{ij}]$ and $\forall s^f_j \in S^f,\ \forall G_g \subset G$: */
if $\left(\sum_{u^g_i \in U} x_{ij} > 0\right)$ **then**
  $b_{gj} \leftarrow 1$
else
  $b_{gj} \leftarrow 0$
end

home service providers [26]–[28]. To solve this problem, we relax the ILP optimization problem into a linear programming (LP) formulation by relaxing the integer decision variables, deriving analytical and solver-based solutions for the relaxed decision variables, and then restoring the integer decision variables using a greedy rounding technique (depicted in Algorithm 1) for reaching the integer values of the decision variables. In 1), we sort the decision variables in descending order and round each value. In the case of the rounded value equals 1, we make sure that the value of rounding achieves constraints; thus, the decision variable x_{ij} is set to 1, otherwise, it is set to zero. After that, we determine the value of b_{gj} based on the condition $\left(\sum_{u^g_i \in U} x_{ij} > 0\right)$. When the previous condition is true, b_{gj} is set to one. Otherwise, it is set to 0.

There are several mathematical methods to solve multi-objective problems [26]. We use the weight-sum method [27] to convert the multi-objective problem into a scalar optimization problem in which the objective function is a weighted sum of the components of the original multi-objective function, as given by Eq. (8), where $0 \leq w \leq 1$ is a scalar weighting factor associated with the objective function for all home service providers, and $T^{resp}_{ijg,norm}$ is the normalized response delay of task $\mathcal{T}_{ig}$ at Service Provider s_j, and is given by Eq. (9).

$$\min_{x_{ij},\ b_{gj}} \frac{w\ C_{SP}}{C_{SP_{Max}}} + \frac{(1-w)}{n} \sum_{g=1}^{z} \sum_{u^g_i \in U^j} \sum_{s_j \in S} x_{ij} T^{resp}_{ijg,norm} \tag{8}$$

$$S.T. \qquad (5b - 5j)$$

$$T^{resp}_{ijg,norm} = \frac{T^{resp}_{ijg} - T^{resp}_{Min}}{T^{resp}_{Max} - T^{resp}_{Min}} \tag{9}$$

Note that T^{resp}_{Min}, and T^{resp}_{Max} denote the minimum and maximum average response time in the MEC system. We relax the optimization problem in Eq. (8) into a linear programming (LP) formulation. The objective function and constraints are similar to the integer linear problem in Eq. (5) except constraints (5i) and (5j), which are replaced by $0 \leq x_{ij}$, and $b_{gj} \leq 1$. We derive an analytical solution using the KKT conditions and Lagrangian analysis. The details of the derived solution are presented in the next subsection.

C. Analytical Solution

The linear problem version of Eq. (5) is a convex optimization problem. The KKT conditions provide necessary and sufficient conditions for the optimal solution [29]. Thus, applying the KKT conditions to the constraints of the problem, and the gradient of the Lagrangian function facilitate finding the analytical solution of the decision variables x_{ij}, and b_{gj}. The Lagrangian function associated with the linear optimization problem is given by Eq. (13) in Appendix A, where X is the vector of offloading decisions (i.e., $X = [x_{ij}]$), B is the vector of placement replication decisions (i.e., $B = [b_{gj}]$, $\forall u^g_i \in U,\ \forall s_j \in S,\ \forall G_g \subset G$, and:

- $\lambda_{home} = [\lambda_{home_j}]$, $\lambda_{foreign} = [\lambda_{foreign_j}]$, $\alpha = [\alpha_i]$, $\sigma_1 = [\sigma_{1_{ij}}]$, $\sigma_2 = [\sigma_{2_{ij}}]$, $\theta = [\theta_{gil}]$, are the associated Lagrange multiplier to inequality constraints 5c, 5d, 5e, 5f, 5g, 5h, respectively.
- $\beta = [\beta_{ij}]$, $\gamma = [\gamma_{ij}]$ are the associated Lagrange multiplier to the lower bound and upper bound inequality constraints of x_{ij}, respectively.
- $\mu = [\mu_{gj}]$, $\eta = [\eta_{gj}]$ are the associated Lagrange multiplier to the lower bound and upper bound inequality constraints of b_{gj}, respectively.
- $\nu = [\nu_i]$ is the associated Lagrange multiplier to the equality constraint 5b.

Appendix B shows the derived equations by applying the KKT conditions. After applying the KKT conditions on the equality and inequality constraints, the following theorem illustrates the optimal solution of the linear problem.

Theorem 1: The optimal offloading decision of the linear optimization problem can be expressed as shown in Eq. (22), Eq. (23), and Eq. (26).

Proof: The proof of Theorem 1 is in Appendix C.

It can be deduced from Eq. (22), Eq. (23), and Eq. (26) that the analytical solutions do not yield closed-form expressions for the relaxed problem. Thus, we solve the linear problem using a numerical solver. We then employ the greedy rounding algorithm, described in Algorithm 1, to restore the binary values of the decision variables x_{ij} and b_{gj}. This algorithm is based on the one defined in [30].The worst-case time

complexity of the greedy rounding algorithm is $O(n^2k^2)$, which is classified under polynomial complexity. Note that the time complexity of the linear relaxation-based solution is the summation of the complexities of solving the linear problem and the rounding algorithm. This time complexity has been shown to be less than that of solving the original ILP [31]–[33]. Consequently, we focus on solving the linear relaxation-based (LR) problem for the proposed method.

IV. SRMSP-Distributed Allocation (SRMSP-DA)

In this section, we solve the aforementioned resource allocation problem in the context of distributed scenarios. We present the service replication game (SRG) and use a game-theoretic approach to solve the problem in a distributed manner. The distributed allocation approach can reduce the burden of finding a centralized optimal solution by making allocation decisions for home service providers individually, while achieving a collectively satisfactory allocation solution. It also scales with the size of the SRG problem. In addition, compared to the centralized approach, a distributed game-theoretic approach can find a solution more quickly, which fulfills service needs for low latency in the MEC environment [34]. We provide a detailed discussion of the employed game-theoretic approach below.

A. Game Formulation

The SRG game is built to find task allocation and service replication indicators for the home service provider(s) that offload their local users' tasks to the MEC server of the home service provider and the MEC servers of foreign service providers in a cost-efficient manner. In the SRG game, the home service providers are players who decide $[X, B]$. Let $X_{-j} = [X_1, \dots X_{j-1}, X_{j+1}, \dots, X_m]$ and $B_{-j} = [B_1, \dots, B_{j-1}, B_{j+1}, \dots, B_m]$ be used to represent all home service providers allocation decisions except home service provider j. Given the other home service providers' decisions X_{-j}, home service provider j strives to make a decision X_j to minimize its utility function, which is defined as given by Eq. (10), while achieving the user group's QoS requirements. Note that $T^{resp}_{j,norm}$ is defined as given by Eq. (11) which represents the normalized average response time of the home service provider s^h_j.

$$\min\ UF_j = \min\left(\frac{w\ C_{SP_j}}{C_{SP_{Max_j}}} + (1-w)T^{resp}_{j,norm}\right) \quad (10)$$

$$T^{resp}_{j,norm} = \frac{1}{n_j}\sum_{g=1}^{z}\sum_{u^g_i \in U^j}\sum_{s_j \in S} x_{ij}T^{resp}_{ijg,norm} \quad (11)$$

Based on Eq. (10), the SRG problem can be formulated as a game $\Gamma = (m, \{X_j, B_j\}_{s^h_j \in S^h}, \{UF_j\}_{s^h_j \in S^h})$. In this game, there might be conflicts among home service providers. This is because the users, as well as service allocation to a particular foreign service provider, might prevent other home service providers from being allocated to the same foreign service provider or other foreign service providers. To prevent conflicts among home service providers, we study whether the game admits at least one Nash equilibrium [34].

B. Nash Equilibrium

For the SRG game Γ, if the strategy profile $X^* = [X^*_1, \dots, X^*_m]$ and $B^* = [B^*_1, \dots, B^*_m]$ are Nash equilibria, any strategy x^*_j, B^*_j is optimal for the strategic combination of other home service providers, i.e.,

$$UF_j(X^*_j, B^*_j, X^*_{-j}, B^*_{-j}) \leq UF_j(X_j, B_j, X^*_{-j}, B^*_{-j}).$$

The above definition implies that no home service provider can further increase its utility by unilaterally changing its strategy at equilibrium. Next, we introduce the concept of a potential game to solve the problem [**?**].

C. Potential Game

For the strategy profile (X, B), there exists a function Φ if it meets the criteria given by Eq. (12). We provide proof that the SRG game is an exact potential game in Appendix D.

$$\begin{aligned} &UF_j(X_j, B_j, X_{-j}, B_{-j}) - UF_j(X'_j, B'_j, X_{-j}, B_{-j}) \\ &\quad = \Phi_{(X_{-j}, B_{-j})}(X_j, B_j) - \Phi_{(X_{-j}, B_{-j})}(X'_j, B'_j). \end{aligned} \quad (12)$$

D. Distributed Allocation (DA) Algorithm

Given the set S^h of home service providers and the set S^f of foreign service providers, the SRG game employs an iterative process for home service providers to reach the Nash equilibrium. The worst-case time complexity of the Distributed Allocation Algorithm is $O(mn^2k)$, which falls within the realm of polynomial complexity. As shown in Algorithm 2, the process starts by initializing the allocation by assigning all users to their home service provider $s^h_j \in S^h$. The task allocation process then proceeds, where for each home Service Provider s^h_j, if s^h_j is overloaded, s^h_j looks for a user served by a foreign service provider that has the lowest utility function. Then, s^h_j sends a computing resource request to the detected foreign service provider and waits for an acknowledgment from this foreign service provider. Each foreign Service Provider s^f_j receives multiple requests from all home service providers. If the foreign Service Provider s^f_j can serve the users' tasks, s^f_j sends a positive acknowledgment to the corresponding home service provider. However, if these requests cause s^f_j to be overloaded, s^f_j will select between the users' tasks based on one of two different selection criteria; (1) select the tasks that require more computing resources in order to Maximize the Profit (DA-MP), or (2) Randomly Select (DA-RS) from the received requests from the home service providers. In this case, s^f_j sends a negative acknowledgment to the home Service Provider s^h_j of the rejected tasks.

In case of receiving a positive acknowledgment, the home service provider releases the user to the selected foreign service provider.

Alternatively, if the home service provider receives a negative acknowledgment, this means that the foreign service

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3328595

Algorithm 2: Distributed Allocation Algorithm

Input : $m,\ k,\ T_{ijg}^{resp},\ \tau_j^C,\ \tau_j^T,\ \tau_g^R,\ w,\ \mathcal{T}_{ig},\ G_g,\ G,\ U^j,\ S^h,\ S^f,\ U$
Output: X_j and $B_j,\ \forall s_j^h \in S^h$
Initialization:
$x_{ij} \leftarrow 1,\ \forall u_i^g \in U^j,\ \forall G_g \subset G,\ \forall s_j^h \in S^h$
Task Allocation Process:
for $s_j^h \in S^h$ **do**
  while $\left(\sum_{u_i^g \in U^j} f_{ig} x_{ij} > C_j^h\right)$ **do**
    for $s_j^f \in S^f$ **do**
      for $u_i^g \in U^j$ **do**
        if *constraints (5b-5j) are binding* **then**
          Compute UF_j using (10)
        else
          $UF_j \leftarrow 0$
        end
      end
    end
    Determine s_j^f, which has $min(UF_j)$, $x_{ij} = 1$, and $min(UF_j) \neq 0$
    Send a request to that s_j^f `/* asking for computing resources */`
    Wait until s_j^h receives an acknowledgment from s_j^f
    if s_j^h *receives a positive acknowledgment* **then**
      $x_{ij} \leftarrow 1,\ s_j^f \in S^f$
      $x_{ui} \leftarrow 0,\ s_j^h \in S^h$
    else
      `/* ` s_j^h `receives a negative acknowledgment */`
      $UF_j \leftarrow 0$
    end
  end
end
Replication Allocation Process: `/* Based on the task allocation process */`
if $\left(\sum_{u_i^g \in U^j} x_{ij} > 0\right)$ **then**
  $b_{gj} \leftarrow 1$
else
  $b_{gj} \leftarrow 0$
end

provider will be overloaded if it accepts this request. In this case, we remove the foreign service provider from future allocations by setting the users' utility functions at this foreign service provider to zero. Finally, we check if the home service provider is still overloaded and repeat the process if it is overloaded. In the replication allocation process and after allocating all tasks, we detect the replication indicator b_{gj} for Group $G_g \in G$ at the foreign Service Provider $s_j^f \in S^f$ in the replication allocation process. If at least one user is located at s_j^f, b_{gj} is set to 1. Otherwise, it is set to 0.

V. Performance Evaluation

In this section, we evaluate the performance of SRMSP and SRMSP-DA compared to a baseline task replication scheme that minimizes the average response delay but disregards the operational cost of service providers [18]. This baseline scheme is referred to as the Delay-Oriented Service Replication (DOSR) scheme. In contrast to SRMSP, DOSR's objective is to minimize the average response delay without considering the operational cost of service providers. In addition, we compare SRMSP and SRMSP-DA to the Distance-Based Greedy (DBG) scheme [7]. In DBG, users close to their home service provider offload their tasks to the latter, whereas the remaining users offload their tasks to the closest available foreign service provider (i.e., the closest one that can still satisfy their group's QoS requirements). Note that SRMSP-DA is implemented twice, once using the Random Selection approach (SRMSP-DA-RS), and once using the Maximum Profit approach (SRMSP-DA-MP). For simplicity, we refer to SRMSP-DA-RS and SRMSP-DA-MP as DA-RS and DA-MP, respectively. We use the following performance metrics: 1) the average response delay, which is calculated as the average transmission, computation, replication, and roaming time over all users in all edge servers, 2) the operational cost, which is calculated as the sum of the total fees paid by home service providers to foreign service providers to serve their users, and 3) the average number of iterations it takes for all home service providers to reach the Nash equilibrium.

A. Simulation Setup

We implement SRMSP, DA-MP, DA-RS, and DOSR using MATLAB [35]. Eight service providers are deployed in an area of approximately 1000 m x 1000 m. The number of overloaded home service providers and underloaded foreign service providers is set to 4. The data rate is uniformly distributed in the range of [30-50] bytes/sec. The computation frequency of the edge servers follows a uniform distribution in the range of [400-800]. The data size of the tasks follows a uniform distribution in the range of [0.01-0.05]. The workload intensity of the tasks follows a random distribution in the range of [100-700]. The price of the communication rate τ_j^T and the price of computation resource τ_j^C of service provider s_j, follow a uniform distribution in the range of [0.01-0.05] and [0.1-1], respectively. The unit cost of replicating the essential service of Group G_g, τ_g^R, follows a uniform distribution in the range of [1-10].

Edge computing service providers are placed at fixed locations in the simulation area. All the groups of user devices are randomly deployed within 200 m of their home service providers. There are up to three groups of users, each group plays a multi-player game using a virtual reality application with different QoS requirements. The maximum delay difference between any two users (belonging to the same group) follows a uniform distribution in the range of [10-15], and the maximum delay of the group is selected randomly in the range of [25-40]. Unless otherwise specified, the number of users is set to 200. The weight coefficient w is set to 0.7.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3328595

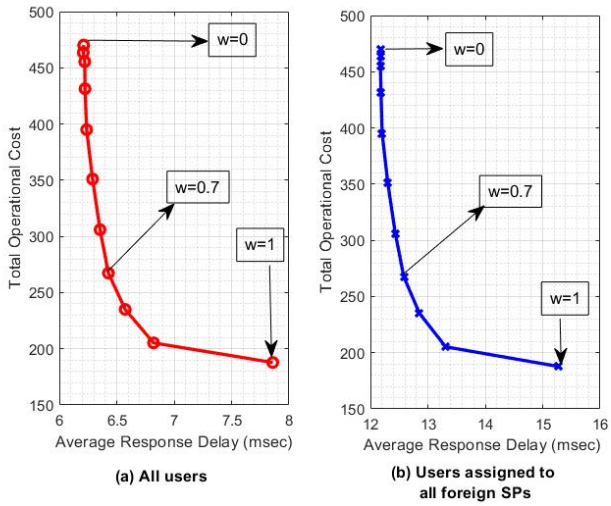

Fig. 2: Pareto optimality of SRMSP

B. Results and Discussion

In our experiments, we evaluate the performance of DOSR, SRMSP, DA-RS, and DA-MP over a varying number of users to assess their scalability. In addition, we present the trade-off (i.e., Pareto optimal) between the two objectives in SRMSP; the total cost and the average response delay. We repeated the experiments 20 times for each instance, and we then determined the confidence intervals using the t-distribution. According to that, we are 95% confident that the results are within ± 10 of the mean results reported.

1) Pareto Optimal Results

Figure 2 shows the trade-off (i.e., Pareto optimal) between the two objectives in SRMSP when the weight coefficient w varies from 0 to 1. Note that the average response time is considered for all users in Figure 2(a), and for the subset of users whose tasks are assigned to foreign service providers in Figure 2(b). If w=1, the optimum solution is reached by minimizing the total operational cost of service providers while disregarding the average response delay, and vice versa if w=0. Thus, it can be observed that when w=1, SRMSP yields the lowest total operational cost and the highest average response delay. In contrast, SRMSP yields the highest operational cost and lowest average response delay when w=0. Note that the difference in the range of the average response time between Figure 2(a) and Figure 2(b) is attributed to the fact that all users execute their tasks at their home service providers unless overloaded, thus leading to lower average response delay in Figure 2(a). The remaining tasks of unserved users are assigned to foreign service providers. The total amount invoiced to home service providers (to cover the foreign service provider's expenses) is the same in Figure 2(a) and Figure 2(b), since the value depends on the assignment of tasks to foreign service providers only.

It is worth mentioning that deploying a service at the foreign service providers' edge servers and reusing it can lead to reduced latency and operational costs, thus enhancing the scalability of real-time services by making them available on-demand at underloaded edge servers of foreign service providers. This way, all users can be served concurrently, especially in situations when home service providers are unable to handle the load from all user groups effectively [36], [37].

2) Impact of the Number of Users

Figure 3(a) illustrates the total operational cost of four schemes: DOSR, SRMSP, DA-RS, DA-MP, and DBG, across different user counts. Notably, the operational cost becomes evident for all schemes when the number of users reaches 120. This suggests that home service providers can independently serve all users when their count is below 120. However, as the number of users increases, the total cost rises in all schemes due to increasing overload on home service providers. Consequently, there is a greater need for task replication at underloaded foreign service providers to accommodate unserved users and enhance system scalability. However, this results in increased costs for home service providers who have to pay for the services rendered by foreign service providers. As depicted in Figure 3(a), it can be observed that SRMSP, DA-RS, and DA-MR outperform DOSR in terms of the total operational cost, achieving significant improvements of up to 50%, 37%, and 37%, respectively. This is because DOSR focuses solely on minimizing the average response delay and overlooks the operational costs incurred by service providers, whereas SRMSP, DA-RS, and DA-MR account for both aspects. It can also be observed that SRMSP, DA-RS, and DA-MR outperform DBG, achieving significant improvements of up to 54%, 43%, and 43%, respectively. This can be attributed to the fact that DBG concentrates solely on offloading tasks to the closest available foreign service provider without considering the operational costs. In contrast, SRMSP, DA-RS, and DA-MR successfully strike a balance by enhancing system scalability while minimizing operational costs. It is worth noting that DA-RS and DA-MR yield nearly identical results, approaching the optimal solution provided by SRMSP with a gap of up to 20%.

Figure 3(b) depicts the average response delay of DOSR, SRMSP, DA-RS, DA-MP, and DBG over varying numbers of users. As the user count increases, there is a noticeable rise in the average response delay for all schemes. This escalation can be attributed to the increased likelihood of home service providers becoming overwhelmed. Consequently, there is a higher frequency of service replication for user requests at foreign service providers, leading to delays in transmission, roaming, and replication, thus contributing to an overall increase in the average response delay. DOSR achieves the minimum response delay among all the schemes since it focuses solely on optimizing the delay. Despite the trade-off between the operational cost and response delay in SMRSP, DA-RS, and DA-MP, they still manage to closely approach DOSR, with a small gap of up to 4%, 7%, and 7%, respectively. The reason for this comparable performance is that SRMSP, DA-RS, and DA-MP strive to optimize both the operational cost and the average response delay. Considering that more weight is given to optimizing the operational cost (w=0.7), SRMSP

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3328595

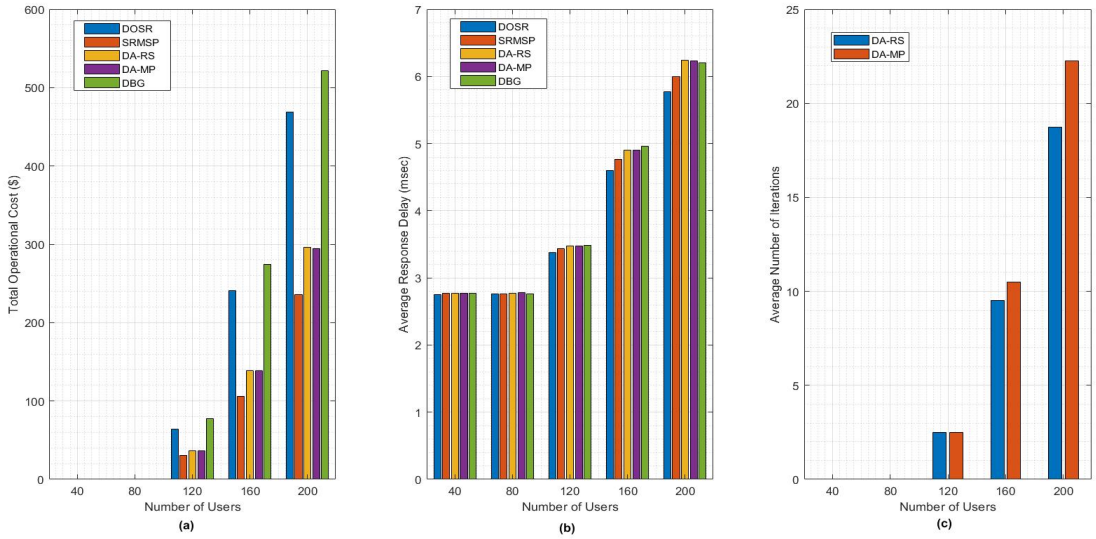

Fig. 3: Performance results of DOSR, SRMSP, DA-RS, and DA-MP over varying number of users.

significantly outperforms DOSR by up to 50% in terms of operational cost, while still managing to yield a small gap of up to 4% in terms of response delay. Additionally, DA-RS and DA-MP significantly outperform DOSR by up to 37% and 37%, respectively, in terms of operational cost, while still rendering a small gap of up to 7% and 7%, respectively, in terms of response delay. Furthermore, SRMSP, DA-RS, and DA-MP outperform DBG by a small margin of up to 4%, 0.1%, and 0.2%, respectively, because offloading tasks to the closest available foreign service provider reduces the transmission delay and, therefore, the average response delay.

Figure 3(c) shows the average number of iterations rendered by DA-RS and DA-MP over varying number of users. It is worth mentioning that we exclusively focus on presenting the number of iterations for the two distributed solution variants, DA-RS and DA-MP, as they rely on heuristic approaches. Conversely, DOSR, SRMSP, and DBG are centered around optimization problems that do not involve iterative processes; they are resolved by employing a numerical solver. Intuitively, as the number of users increases, the number of iterations increases in both DA-RS and DA-MP. As the number of users increases, home SPs take more time to determine the suitable foreign SPs at which home service providers can replicate their tasks. It can be noted that DA-RS takes fewer iterations than DA-MP, with an improvement of up to 15%. This can be attributed to the fact that foreign SPs in DA-MP take more time to respond to home SPs in order to reach their goal, which is maximizing their profit. In contrast, foreign SPs rely on random selection and does not involve any criteria that need to be optimized.

VI. Conclusion

In this paper, we have proposed the Service Replication between Multiple Service Providers (SRMSP) scheme. SRMSP enables service scalability at overloaded home service providers by fostering cost and delay-efficient service replication at underutilized computational resources of foreign service providers. The service replication problem is formulated as an Integer Linear Program (ILP), and an analytical solution is derived using the KKT conditions and Lagrangian analysis. In addition, we have proposed the SRMSP-Distributed Allocation (SRMSP-DA) scheme. SRMSP-DA solves the problem in distributed scenarios, and in a time-efficient manner using a game-theoretic strategy. We have proposed two versions of SRMSP-DA; one that considers the Maximum Profit (DA-MP) of foreign service providers during the selection process, and one that applies Random Selection (DA-RS). Extensive simulations have shown that DA-RS performs the same as DA-MP, while taking 15% fewer iterations. In addition, it has been shown that SRMSP and SRMSP-DA yield significant improvements of up to 50% and 37%, respectively, in terms of the total operational cost compared to a baseline scheme that focuses on the delay only without accounting for the operational cost. Such improvements are achieved while closely approaching the latter in terms of the average response delay, yielding a small gap of up to 4% and 7%, respectively. Moreover, extensive simulations have demonstrated that SRMSP and SRMSP-DA exhibit significant operational cost reductions, of up to 54% and 43%, respectively, compared to the DBG scheme. These improvements are achieved while maintaining average response delay levels that are nearly on bar with DBG, with only a small gap of up to 4% and 0.2%, respectively. In addition, SRMSP-DA yields a small gap of up to 20% and 4% in terms of the total operational cost and average response delay, respectively, compared to the optimal solution rendered by SRMSP. In the future, we plan to further improve the delay by predicting the load at home service providers and using such predictions to enable proactive service replication.

VII. Acknowledgment

This research is supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC)

under grant number ALLRP 549919-20, and a grant from Distributive, Ltd., Kingston, ON, Canada

REFERENCES

[1] C. M. Diaz, K. K. Choo, and A. Zunino, " Sharpening the edge: Towards improved edge computing environment for mobile and IoT applications," *Future Generation Computer Systems*, vol. 107, June 2020, pp. 1130–1133.

[2] R. Singh, R. Sukapuram and S. Chakraborty, "Mobility-aware Multi-Access Edge Computing for Multiplayer Augmented and Virtual Reality Gaming," *in proc. of IEEE 21st International Symposium on Network Computing and Applications (NCA)*, Boston, MA, USA, 2022, pp. 191–200, doi: 10.1109/NCA57778.2022.10013599.

[3] X. Deng, J. Yin, P. Guan, N. N. Xiong, L. Zhang and S. Mumtaz, "Intelligent Delay-Aware Partial Computing Task Offloading for Multiuser Industrial Internet of Things Through Edge Computing," *in IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 2954–2966, Feb. 2023, doi: 10.1109/JIOT.2021.3123406.

[4] K. Cao, Y. Liu, G. Meng, and Q. Sun, "An overview on edge computing research," *IEEE access*, vol. 8, May 2020, pp. 85714–85728.

[5] Z. Sharif, L. T. Jung, I. Razzak and M. Alazab, "Adaptive and Priority-Based Resource Allocation for Efficient Resources Utilization in Mobile-Edge Computing," *in IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3079–3093, Feb. 2023, doi: 10.1109/JIOT.2021.3111838.

[6] A. Ahmed and E. Ahmed, "A Survey on Mobile Edge Computing," *in proc. of IEEE 10th International Conference on Intelligent Systems and Control (ISCO)*, 2016, pp. 1–8.

[7] W. Du, T. Lei, Q. He, W. Liu, Q. Lei, H. Zhao, and W. Wang, "Service Capacity Enhanced Task Offloading and Resource Allocation in Multi-Server Edge Computing Environment," Mar. 2019, *arXiv preprint arXiv:1903.04709*.

[8] Y. Zhang, X. Chen, Y. Chen, Z. Li, and J. Huang, "Cost Efficient Scheduling for Delay-Sensitive Tasks in Edge Computing System," *in proc. of IEEE International Conference on Services Computing (SCC)*, Jul. 2018, pp. 73–80.

[9] S. Kumar, R. Gupta, K. Lakshmanan, and V. Maurya, "A game-theoretic approach for increasing resource utilization in edge computing enabled internet of things", *IEEE Access*, vol. 10, May 2022, pp. 57974–57989.

[10] B. Shabir, A. W. Malik, A. U. Rahman, M. A. Khan, Z. Anwar, "A Reliable Learning Based Task Offloading Framework for Vehicular Edge Computing", *in proc. of 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, May 2022, IEEE, pp. 1–6.

[11] Y. Zhu, B. Mao, and N. Kato, "A dynamic task scheduling strategy for multi-access edge computing in IRS-aided vehicular networks", *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 4, Mar. 2022, pp. 1761–1771.

[12] K. T. Rodrigues, and N. Kato, "Hybrid Centralized and Distributed Learning for MEC-Equipped Satellite 6G Networks", *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 4, Feb. 2023, pp. 1201–1211.

[13] S. Verma, Y. Kawamoto, and N. Kato, "Energy-efficient group paging mechanism for QoS constrained mobile IoT devices over LTE-A pro networks under 5G," *IEEE Internet of Things Journal*, vol. 6, no.5, Jul. 2019, pp. 9187–9199.

[14] H. Guo, and J. Liu, "UAV-enhanced intelligent offloading for Internet of Things at the edge", *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, Nov. 2021, pp. 2737–2746.

[15] H. Guo, and J. Liu, "Collaborative computation offloading for multi-access edge computing over fiber–wireless networks", IEEE Transactions on Vehicular Technology, vol. 67, no. 5, Jan. 2018, pp. 4514–4526.

[16] A. M. Zaki, S. Sorour, "Proactive Migration for Dynamic Computation Load in Edge Computing," *in proc. of ICC IEEE International Conference on Communications*, May 2022, pp. 4275–4280.

[17] S. A. Mohamed, S. Sorour, and H. S. Hassanein, "Group-Delay Aware Task Offloading with Service Replication for Scalable Mobile Edge Computing," *in proc. of IEEE Global Communications Conference (GLOBECOM)*, Dec. 2020, pp. 1–6.

[18] S. A. Mohamed, S. Sorour, and H. S. Hassanein, "Group Delay-Aware Scalable Mobile Edge Computing Using Service Replication, " *IEEE Transactions on Vehicular Technology,*, vol. 71, no. 11, Nov. 2022, pp. 11911–11920.

[19] X. Ma, S. Zhang, W. Li, P. Zhang, C. Lin, and X. Shen, "Cost-Efficient Workload Scheduling in Cloud Assisted Mobile Edge Computing," *in proc. of IEEE/ACM 25th International Symposium on Quality of Service (IWQoS)*, June. 2017, pp. 1–10.

[20] J. L. Herrera, J. Galán-Jimnez, J. Garcia-Alonso, J. Berrocal, and J. M. Murillo," Joint Optimization of Response Time and Deployment Cost in Next-Gen IoT Applications", *IEEE Internet of Things Journal*, vol. 10, no. 5, April 2022, pp. 3968–3981.

[21] J. Liu, C. Liu, B. Wang, G. Gao, and S. Wang, "Optimized task allocation for IoT application in mobile-edge computing," *IEEE Internet of Things Journal*, vol. 9, no.13, June 2021, pp. 10370–10381.

[22] J. Kim, L. M. Contreras, P. Greto, H. Magnusson, H. Woesner, D. Fritzsche, L. Cominardi, C. J. Bernardos, " GiLAN Roaming: Roam Like at Home in a Multi-Provider NFV Environment", *In International Symposium on Networks, Computers and Communications (ISNCC), IEEE*, June 2018, pp. 1–4.

[23] S. Song, and J. M. Chung, "Sliced NFV service chaining in mobile edge clouds," *in proc. of IEEE 19th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp. 292–294, September 2017.

[24] L. A. Haibeh, M. C. Yagoub, and A. Jarray, "Agile Design and Dimensioning of MEC-NFV Infrastructure to Support Heterogeneous Service Chains," *in proc of IEEE Global Communications Conference (GLOBECOM)*, pp. 3670–3675, December 2022.

[25] J. Park, S. Solanki, S. Baek, and I. Lee, "Latency minimization for wireless powered mobile edge computing networks with nonlinear rectifiers," *IEEE Transactions on Vehicular Technology,* vol.70, no. 8, June 2021 pp. 8320–8324.

[26] G. W. Evans,"An Overview of Techniques for Solving Multiobjective Mathematical Programs," *Management Science*, vol. 30, no. 11, Nov. 1984, pp. 1268–1282.

[27] M. Ehrgott, "Multicriteria Optimization," *Springer Science and Business Media*, Vol. 491, May 2005.

[28] S. S. Rao, "Multiobjective Optimization in Structural Design with Uncertain Parameters and Stochastic Processes," *Aiaa Journal*, vol. 22, no. 11, Nov. 1984, pp. 1670–1678.

[29] S. Boyd and L. Vandenberghe, "Convex Optimization," *Cambridge Univ. Press*, Mar. 2004.

[30] R. Cont, and M. Heidari, "Optimal Rounding under Integer Constraints," *arXiv preprint arXiv:1501.00014*, Dec. 2014.

[31] C. H. Papadimitriou, "On The Complexity of Integer Programming," *Journal of the ACM (JACM)*, vol. 28, no. 4, 1981, pp. 765–768.

[32] N. Megiddo, "On The Complexity of Linear Programming," *In Advances in economic theory: Fifth world congress. London: Cambridge University Pres*, Jan. 1987, pp. 225–268.

[33] C. C. Gonzaga, "On The Complexity of Linear Programming," *Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo,* vol. 2, no. 2, 1995, pp 197–207.

[34] Q. He, G. Cui, X. Zhang, F. Chen, S. Deng, H. Jin, Y. Li, and Y. Yang, "A game-theoretical approach for user allocation in edge computing environment," *IEEE Transactions on Parallel and Distributed Systems,* vol. 31, no. 3, Sept. 2019, pp. 515–529.

[35] [Online] https://www.mathworks.com/products/optimization.html.

[36] L. Zhao, W. Tan, B. Li, Q. He, L. Huang, Y. Sun, and Y. Yang, "Joint shareability and interference for multiple edge application deployment in mobile-edge computing environment," *IEEE Internet of Things Journal*, vol. 9, no. 3, June 2021, pp. 1762–1774.

[37] P. Wang, J. Xu, M. Zhou, and A. Albeshri, "Budget-constrained Optimal Deployment of Redundant Services in Edge Computing Environment," *IEEE Internet of Things Journal*, Jan 2023.

[38] J. R. Marden, G. Arslan, and J. S. Shamma, "Cooperative Control and Potential Games," *IEEE Transactions on Systems Man, and Cybernetics, Part B (Cybernetics),* vol. 39, no. 6, Apr. 2009, pp. 1393–1407.

Shimaa A. Mohamed received her Ph.D. degree from the School of Computing, Queen's University, Canada, in 2021. She is a Postdoctoral Fellow at the School of Computing, Queen's University. She received the BSc and the MSc degrees in electrical and communication engineering from Alexandria University, Egypt, in 2010 and 2015, respectively. She is a researcher at the Informatics Research Institute (IRI), City of Scientific Research and Technological Applications (SRTA-city). Her research interests include edge computing, cloud computing, the Internet of Things, software-defined networking, and sensor networks.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3328595

Sameh Sorour (SM IEEE), 1979-2021, was an assistant professor at the School of Computing, Queen's University, where he led several projects on autonomous and connected vehicles, edge intelligence and wireless networks and services. He was a senior IEEE member and an Editor for IEEE Communications Letters. His research and educational interests lie in the broad areas of advanced computing, learning, and networking technologies for cyber-physical and autonomous systems. Topics of particular interest include cloud/edge/IoT networking, computing, learning, and their applications in multimodal/coordinated autonomous driving, autonomous/electric mobility on demand systems, and cyber-physical systems.

Sara A. Elsayed Sara A. Elsayed received her Ph.D. degree from the School of Computing, Queen's University, Canada. She is a Postdoctoral Fellow and an Adjunct Assistant Professor at the School of Computing, Queen's University. She manages and oversees a large-scale research project for democratizing edge computing, involving an industrial collaboration with Distributive, Ltd. She has received the 2023 Queen's School of Computing (QSC) research award in recognition of her contributions. Prior to joining Queen's University, Elsayed was an Assistant Lecturer in the Faculty of Computers and Information, Cairo University, Egypt. She was the Co-founder and Program Manager of the Teaching Assistant Qualifying Program (TIC), as well as an Assistant Focal Point in the Support Office for Research Cooperation & Mobility (SORCAM), Engineering Sector, Cairo University. She is a certified Cisco Networking Academy Instructor. Her research interests include edge computing, vehicular networks, caching in VANETs, information-centric networks, and Internet of Things. She has several publications in top venues and is a member of the IEEE.

Hossam S. Hassanein is a leading researcher in the areas of broadband, wireless and mobile networks architecture, protocols, control and performance evaluation. His record spans more than 600 publications in journals, conferences and book chapters, in addition to numerous keynotes and plenary talks in flagship venues. In recognition of his significant contributions, Hassanein has received several awards, including the 2003 Ontario Champions of Innovations Award, 2016 IEEE Communications Society Communications Software Committee Technical Achievement Award and the 2020 IEEE IoT, Ad Hoc, and Sensor Networks Committee Technical Achievement and Recognition Award. He is also the recipient of the 2015 Excellence in Graduate Students Supervision at Queen's University. Hassanein is the Former Chair of the IEEE Communication Society Technical Committee on IoT, AdHoc, and Sensor Networks. He is a Fellow of the IEEE and an IEEE Communications Society Distinguished Speaker.

Appendix A
Lagrangian Function

$$
\begin{aligned}
&L(X, B, \lambda_{home}, \lambda_{foreign}, \alpha, \theta, \sigma_1, \sigma_2, \beta, \gamma, \mu, \eta, \nu) = \\
&\frac{w}{C_{SP_{Max}}}\left(\sum_{s_j \in S^f}\sum_{u_i \in U}(C_j^S * x_{ij}) + \sum_{s_j \in S^f}\sum_{g=1}^{z}(b_{gj} * uc_{G_g}^R)\right) \\
&+ \frac{(1-w)}{n}\sum_{s_j \in S}\sum_{g=1}^{z}\sum_{u_i^g \in U^j}\left(T_{ijg,norm}^{resp} x_{ij}\right) \\
&+ \sum_{s_j \in S^h}\lambda_{home_j}\left(\sum_{u_i^g \in U^j}(f_{ig}x_{ij}) - F_{s_i^h}^h\right) \\
&+ \sum_{s_j \in S^f}\lambda_{foreign_j}\left(\sum_{u_i \in U}(f_{ig}x_{ij}) - C_j^f\right) \\
&+ \sum_{g=1}^{z}\sum_{u_i^g \in U^j}\alpha_i\left(\sum_{s_j \in S}(T_{ijg}^{resp}x_{ij}) - T_g^{max}\right) \\
&+ \sum_{g=1}^{z}\sum_{u_i^g \in U^j}\sum_{u_l^g \in U^j, i \neq l}\theta_{gil}\left((\sum_{s_j \in S}x_{ij}T_{ijg}^{resp}) - (\sum_{s_j \in S}x_{lj}T_{ljg}^{resp}) - d_g\right) \\
&+ \sum_{s_j \in S^f}\sum_{g=1}^{z}\sigma_{1_{gj}}\left(\frac{\sum_{u_i^g \in U^j}x_{ij}}{y_g} - b_{gj}\right) \\
&+ \sum_{s_j \in S^f}\sum_{g=1}^{z}\sigma_{2_{gj}}\left(b_{gj} - \sum_{u_i^g \in U^j}x_{ij}\right) - \sum_{u_i \in U}\sum_{s_j \in S}\beta_{ij}x_{ij} \\
&+ \sum_{u_i \in U}\sum_{s_j \in S}\gamma_{ij}(x_{ij} - 1) - \sum_{s_j \in S^f}\sum_{g=1}^{z}\mu_{gj}b_{gj} \\
&+ \sum_{s_j \in S^f}\sum_{g=1}^{z}\eta_{gj}(b_{gj} - 1) + \sum_{u_i \in U}\nu_i\left(\sum_{s_j \in S}x_{ij} - 1\right) \\
&\forall s_j^h \in S^h,\ \forall s_j^f \in S^f,\ \forall G_g \subset G,\ \forall u_i^g, \in U
\end{aligned}
\tag{13}
$$

Appendix B
KKT Conditions

As shown below, by applying the KKT conditions (i.e., dual feasibility, stationarity, primal feasibility, and complementary slackness) on the equality and inequality constraints of the linear problem in Eq. (8), we derive Eq. (14) - Eq. (17j).

- Dual feasibility
  This ensures that the dual variables corresponding to the inequality constraints are non-negative.

$$
\begin{aligned}
&\lambda_{home_j}^*,\ \lambda_{foreign_j}^*,\ \alpha_i^*,\ \theta_{gil}^*,\ \sigma_{1_{gj}}^*,\ \sigma_{2_{gj}}^*,\ \beta_{ij}^*, \\
&\gamma_{ij}^*,\ \mu_{gj}^*\ and\ \eta_{gj}^* \geq 0, \\
&\qquad \forall s_j \in S,\ \forall G_g \subset G,\ \forall u_i^g \in U
\end{aligned}
\tag{14}
$$

- Stationarity
  It ensures that the gradient of the Lagrange function in Eq. (10) with respect to the decision variables x_{ij}, x_{lj}, and b_{gj} is zero at the optimal solutions, which indicates

that the solutions x^*_{ij}, x^*_{lj}, and b^*_{gj} are critical points where the objective function is not changing with respect to the decision variables such that:

$$\frac{\partial L(X^*, B^*, \lambda^*_{home_j}, \ldots, \nu^*)}{\partial x^*_{ij}} = 0,$$

$$\frac{1-w}{n} T^{resp}_{ijg,norm} + \lambda^*_{home_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + \theta^*_{gil} T^{resp}_{ijg} - \beta^*_{ij} + \gamma^*_{ij} + \nu^*_i = 0. \quad \forall s^h_i \in S^h,\ \forall G_g \subset G,\ \forall u^g_i,\ u^g_l \in U^j,\ l \neq i \tag{15a}$$

$$\frac{\partial L(X^*, B^*, \lambda^*_{home_j}, \ldots, \nu^*)}{\partial x^*_{ij}} = 0,$$

$$\frac{w}{C_{SP_{Max}}} C_{s_j} + \frac{1-w}{n} T^{resp}_{ijg,norm} + \lambda^*_{foreign_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + \theta^*_{gil} T^{resp}_{ijg} + \left(\frac{\sigma^*_{1_{gj}}}{y_g}\right) - \sigma^*_{2_{gj}} - \beta^*_{ij} + \gamma^*_{ij} + \nu^*_i = 0, \quad \forall s^f_j \in S^f,\ \forall G_g \subset G,\ \forall u^g_i,\ u^g_l \in G_g,\ l \neq i \tag{15b}$$

$$\frac{\partial L(X^*, B^*, \lambda^*_{home_j}, \ldots, \nu^*)}{\partial x^*_{lj}} = 0,$$

$$-\theta_{gil} T^{resp}_{ljg} = 0 \quad \forall s^h_i \in S^h,\ \forall G_g \subset G,\ \forall u^g_i,\ u^g_l \in U^j,\ l \neq i \tag{15c}$$

$$\frac{\partial L(X^*, B^*, \lambda^*_{home_j}, \ldots, \nu^*)}{\partial b^*_{gj}} = 0,$$

$$\frac{w}{C_{SP_{Max}}} \tau^R_g - \sigma^*_{1_{gj}} + \sigma^*_{2_{gj}} - \mu^*_{gj} + \eta^*_{gj} = 0 \quad \forall s^f_j \in S^f,\ \forall G_g \subset G,\ \forall u^g_i,\ u^g_l \in U^j,\ l \neq i \tag{15d}$$

- Primal feasibility
  It ensures that the decision variables satisfy all the constraints specified in the problem. This includes the equality and inequality constraints of the optimization problem in Eq. (8), as shown below.

$$\sum_{s_j \in S} x^*_{ij} = 1, \qquad \forall u^g_i \in U \tag{16a}$$

$$\sum_{u^g_i \in U^j} x^*_{ij} f_{ig} \leq C^h_j, \qquad \forall s^h_j \in S^h \tag{16b}$$

$$\sum_{u^g_i \in U} x^*_{ij} f_{u^g_i} \leq C^f_j, \qquad \forall s^f_j \in S^f \tag{16c}$$

$$\sum_{s_j \in S} x^*_{ij} T^{resp}_{ijg} \leq T^{max}_g, \qquad \forall G_g \subset G,\ \forall u^g_i \in U^j \tag{16d}$$

$$\frac{\sum_{u^g_i \in U^j} x^*_{ij}}{y_g} - b^*_{gj} \leq 0,\ \forall G_g \subset G,\ \forall s^f_j \in S^f \tag{16e}$$

$$b^*_{gj} - \sum_{u^g_i \in U^j} x^*_{ij} \leq 0, \qquad \forall G_g \subset G,\ \forall s^f_j \in S^f \tag{16f}$$

$$\sum_{s_j \in S} x^*_{ij} T^{resp}_{ijg} - \sum_{s_j \in S} x^*_{lj} T^{resp}_{ljg} \leq d_i, \quad \forall G_g \subset G,\ \forall u^g_i,\ u^g_l \in U^j,\ l \neq i \tag{16g}$$

$$-x^*_{ij} \leq 0, \quad \forall u^g_i \in U,\ \forall s_j \in S \tag{16h}$$

$$x^*_{ij} - 1 \leq 0, \quad \forall u^g_i \in U,\ \forall s_j \in S \tag{16i}$$

$$-b^*_{gj} \leq 0, \quad \forall s^f_j \in S^f,\ \forall G_g \subset G \tag{16j}$$

$$b^*_{gj} - 1 \leq 0, \quad \forall s^f_j \in S^f,\ \forall G_g \subset G \tag{16k}$$

- Complementary slackness
  It involves the multiplication of the Lagrange multiplier with the inequality constraint (5c), (5d), (5e), (5f), (5g), (5h), (5i), and (5j) of the linear problem in Eq. (8), as follows.

$$\lambda^*_{home_j} \left(\sum_{u^g_i \in U^j} x^*_{ij} f_{ig} - F^h_{s^h_i} \right) = 0,\ \forall s^h_i \in S^h \tag{17a}$$

$$\lambda^*_{foreign_j} \left(\sum_{u_i \in U} x^*_{ij} f_{ig} - C^f_j \right) = 0,\ \forall s^f_j \in S^f \tag{17b}$$

$$\alpha^*_i \left(\sum_{s_j \in S} x^*_{ij} T^{resp}_{ijg} - T^{max}_g \right) = 0, \quad \forall G_g \subset G,\ \forall u^g_i \in U^j \tag{17c}$$

$$\sigma^*_{1_{gj}} \left(\frac{\sum_{u^g_i \in U^j} x^*_{ij}}{y_g} - b^*_{gj} \right) = 0,\ \forall G_g \subset G,\ \forall s^f_j \in S^f \tag{17d}$$

$$\sigma^*_{2_{gj}} \left(b^*_{gj} - \sum_{u^g_i \in U^j} x^*_{ij} \right) = 0,\ \forall G_g \subset G,\ \forall s^f_j \in S^f \tag{17e}$$

$$\theta^*_{gil} \left(\sum_{s_j \in S} x^*_{ij} T^{resp}_{ijg} - \sum_{s_j \in S} x^*_{lj} T^{resp}_{ljg} - d_i \right) = 0, \quad \forall G_g \subset G,\ \forall u^g_i,\ u^g_l \in U^j,\ l \neq i \tag{17f}$$

$$-\beta^*_{ij} x^*_{ij} = 0, \qquad \forall u^g_i \in U,\ \forall s_j \in S \tag{17g}$$

$$\gamma^*_{ij} \left(x^*_{ij} - 1 \right) = 0, \qquad \forall\ u^g_i \in U,\ \forall\ s_j \in S \tag{17h}$$

$$-\mu^*_{gj} b^*_{gj} = 0, \qquad \forall s^f_j \in S^f,\ \forall G_g \subset G \tag{17i}$$

$$\eta^*_{gj} \left(b^*_{gj} - 1 \right) = 0, \qquad \forall s^f_j \in S^f,\ \forall G_g \subset G \tag{17j}$$

Appendix C
Proof of Theorem 1

Using the equations derived from the KKT conditions in Appendix B, we can deduce from Eq. (15c) that $T^{resp}_{ljg} \neq 0$ and $\theta^*_{gil} \geq 0$, and thus $\theta^*_{gil} = 0$. By substituting $\theta^*_{gil} = 0$ in Eq. (15a) and Eq. (15b), we get Eq. (18) and Eq. (19), respectively.

$$\frac{1-w}{n}T^{resp}_{ijg,norm} + \lambda^*_{home_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} - \beta^*_{ij} + \gamma^*_{ij} + \nu^*_i = 0.\ \forall s^f_j \in S^f,\ \forall G_g \subset G,\ \forall u^g_i,\ u^g_l \in U^j,\ l \neq i \tag{18}$$

$$\frac{w}{C_{SP_{Max}}} C^S_{s_j} + \frac{1-w}{n}T^{resp}_{ijg,norm} + \lambda^*_{foreign_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + \left(\frac{\sigma^*_{1_{gj}}}{y_g}\right) - \sigma^*_{2_{gj}} - \beta^*_{ij} + \gamma^*_{ij} + \nu^*_u = 0,\ \forall s^f_j \in S^f,\ \forall G_g \subset G,\ \forall u^g_i,\ u^g_l \in U^j,\ l \neq i \tag{19}$$

By multiplying Eq. (18) by x^*_{ij}, and substituting $x^*_{ij}\beta^*_{ij}$ and $x^*_{ij}\gamma^*_{ij}$ in the result of the multiplication with 0 and γ^*_{ij} as in Eq. (17g) and Eq. (17h), respectively, we get Eq. (20).

$$x^*_{ij} = \frac{-\gamma^*_{ij}}{\frac{1-w}{n}T^{resp}_{ijg,norm} + \lambda^*_{home_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + \nu^*_i} = \frac{-\gamma^*_{ij}}{\beta^*_{ij} - \gamma^*_{ij}},\ \forall s^h_j \in S^h,\ \forall u^g_i \in U^i,\ \forall G_g \subset G \tag{20}$$

By multiplying Eq. (19) by x^*_{ij}, and substituting $x^*_{ij}\beta^*_{ij}$ and $x^*_{ij}\gamma^*_{ij}$ in the result of the multiplication with 0 and γ^*_{ij} as in Eq. (17g) and Eq. (17h), respectively, we get Eq. (21).

$$x^*_{ij} = \frac{-\gamma^*_{ij}}{M+F} = \frac{-\gamma^*_{ij}}{\beta^*_{ij} - \gamma^*_{ij}},\ \forall s^f_j \in S^f,\ \forall u^g_i \in U^j,\ \forall G_g \subset G. \tag{21}$$

Note that $M = \frac{w}{C_{SP_{Max}}} C^S_{s_j} + \frac{1-w}{n}T^{resp}_{ijg,norm}$ and $F = \lambda^*_{foreign_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + \left(\frac{\sigma^*_{1_{gj}}}{y_g}\right) - \sigma^*_{2_{gj}} + \nu^*_i$.

From Eq. (20), Eq. (17h), and Eq. (17g), we can deduce that there are only two viable cases where x^*_{ij} has a valid value (0 or 1). These cases are given as follows:

- Case 1: $x^*_{ij} = 0$ when $\gamma^*_{ij} = 0$, and $\beta^*_{ij} > 0$, which means that $\left(\frac{1-w}{n}T^{resp}_{ijg,norm} + \lambda^*_{home_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + \nu^*_i\right) > 0$.
- Case 2: $x^*_{ij} = 1$ when $\gamma^*_{ij} > 0$ and $\beta^*_{ij} = 0$, which means that $\left(\frac{1-w}{n}T^{resp}_{ijg,norm} + \lambda^*_{home_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + \nu^*_i\right) = -\gamma^*_{ij}$.

Otherwise, when $\gamma^*_{ij} = \beta^*_{ij} = 0$, $\gamma^*_{ij} = \beta^*_{ij} > 0$, or $\gamma^*_{ij} > 0$ and $\beta^*_{ij} > 0$, x^*_{ij} has no valid value.

From Eq. (21), Eq. (17h), and Eq. (17g), we can deduce that there are only two viable cases where $x^*_{ij} = 0$ has a valid value (0 or 1). These cases are given as follows:

- Case 1: $x^*_{ij} = 0$ when $\gamma^*_{ij} = 0$, and $\beta^*_{ij} > 0$, which means that $(\frac{w}{C_{SP_{Max}}} C^S_j + \frac{1-w}{n}T^{resp}_{ijg,norm} + \lambda^*_{foreign_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + (\frac{\sigma^*_{1_{gj}}}{y_g}) - \sigma^*_{2_{gj}} + \nu^*_i) > 0$.
- Case 2: $x^*_{ij} = 1$ when $\gamma^*_{ij} > 0$, and $\beta^*_{ij} = 0$, which means that $(\frac{w}{C_{SP_{Max}}} C^S_j + \frac{1-w}{n}T^{resp}_{ijg,norm} + \lambda^*_{foreign_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + (\frac{\sigma^*_{1_{gj}}}{y_g}) - \sigma^*_{2_{gj}} + \nu^*_i) = -\gamma^*_{ij}$.

Otherwise, when $\gamma^*_{ij} = \beta^*_{ij} = 0$ (because of the limitation of x^*_{ij} when γ^*_{ij} and β^*_{ui} tend to zero), $\gamma^*_{ij} = \beta^*_{ij} > 0$, or $\gamma^*_{ij} > 0$ and $\beta^*_{ij} > 0$, x^*_{ij} has no valid value.
Thus, when $\forall s^h_i \in S^h,\ \forall u^g_i \in U^j,\ \forall G_g \subset G$,

$$x^*_{ij} = \begin{cases} 0, & if\ \gamma^*_{ij} = 0,\ \beta^*_{ij} > 0,\ V \neq 0 \\ 1, & if\ \gamma^*_{ij} > 0,\ \beta^*_{ij} = 0,\ V = -\gamma^*_{ij} \end{cases} \tag{22}$$

Note that $V = \frac{1-w}{n}T^{resp}_{ijg,norm} + \lambda^*_{home_j} f_{ig} + \alpha^*_i T^{resp}_{ijg} + \nu^*_i$.
Otherwise, when $\forall s^f_j \in S^f,\ \forall u^g_i \in U,\ \forall G_g \subset G$,

$$x^*_{ij} = \begin{cases} 0, & if\ \gamma^*_{ij} = 0,\ \beta^*_{ij} > 0, \\ & E \neq 0 \\ 1, & if\ \gamma^*_{ij} > 0,\ \beta^*_{ij} = 0, \\ & E = -\gamma^*_{ij} \end{cases} \tag{23}$$

Note that $E = \frac{w}{C_{SP_{Max}}} C^S_j + \frac{1-w}{n}T^{resp}_{ijg,norm} + \lambda^*_{foreign_j} f_{u^g_i} + \alpha^*_i T^{resp}_{ijg} + \left(\frac{\sigma^*_{1_{gj}}}{y_g}\right) - \sigma^*_{2_{gj}} + \nu^*_i$.

As for the other decision variable b^*_{gj}, we multiply Eq. (15d) by b^*_{gj} and substitute $\mu^*_{gj}b^*_{gj}$, and $\eta^*_{gj}b^*_{gj}$ in the result of the multiplication with 0 and η^*_{gj} in Eq. (17i), and Eq. (17j), respectively. Accordingly, Eq. (24) is obtained.

$$b^*_{gj} = \frac{-\eta^*_{gj}}{\frac{w}{C_{SP_{Max}}}\tau^R_g - \sigma^*_{1_{gj}} + \sigma^*_{2_{gj}}} = \frac{-\eta^*_{gj}}{\mu^*_{gj} - \eta^*_{gj}} \tag{24}$$

From Eq. (24) and Eq. (17j), we can deduce that there are only three viable cases where b^*_{gj} has a valid value.
These cases are given as follows:

- Case 1: $b^*_{gj} = 0$ when $\eta^*_{gj} = 0$, which means that $\frac{w}{C_{SP_{Max}}}\tau^R_g - \sigma^*_{1_{gj}} + \sigma^*_{2_{gj}} > 0$.
- Case 2: $b^*_{gj} = 1$ when $\mu^*_{gj} = 0$, which means that $\frac{w}{C_{SP_{Max}}}\tau^R_g - \sigma^*_{1_{s_j i}} + \sigma^*_{2_{gj}} = -\eta^*_{gj}$.
- Case 3: $0 \leq b^*_{gj} \leq 1$, when $\eta^*_{gj} = 0$ and $\mu^*_{gj} = 0$, which means that $\frac{w}{C_{SP_{Max}}}\tau^R_g - \sigma^*_{1_{gj}} + \sigma^*_{2_{gj}} = 0$.
  Thus, there are four cases, based on the values of $\sigma^*_{1_{s_j i}}$ and $\sigma^*_{2_{s_j i}}$, that determine the value of b^*_{gj}. These cases are illustrated as follows:
  - Case 3(a): When $\sigma^*_{1_{gj}} = \sigma^*_{2_{gj}} = 0$, the value of b^*_{gj} in this case invalid because $\left(\frac{w}{C_{SP_{Max}}}\tau^R_g - \sigma^*_{1_{gj}} + \sigma^*_{2_{gj}}\right) \neq 0$
  - Case 3(b): When $\sigma^*_{1_{gj}} = 0$ and $\sigma^*_{2_{gj}} > 0$, the value of b^*_{gj} in this case invalid because $\left(\frac{w}{C_{SP_{Max}}}\tau^R_g - \sigma^*_{1_{gj}} + \sigma^*_{2_{gj}}\right) \neq 0$
  - Case 3(c): When $\sigma^*_{1_{gj}} > 0$ and $\sigma^*_{2_{gj}} = 0$, $b^*_{gj} = \frac{\sum_{u^g_i \in U^j} x^*_{ij}}{y_g}, \quad \forall s^f_j \in S^f,\ \forall G_g \subset G.$
  - Case 3(d): When $\sigma^*_{1_{gj}} > 0$ and $\sigma^*_{2_{gj}} > 0$, the value of b^*_{gj} in this case is invalid because it fails to satisfy the complementary slackness conditions stated in Eq. (17d) and Eq. (17e).

Based on Case 3, we can deduce that b^*_{gj} is given by Eq. (23).

$$b^*_{gj} = \frac{\sum_{u^g_i \in U^j} x^*_{ij}}{y_g}, \qquad \forall s^f_j \in S^f,\ \forall G_g \subset G. \tag{25}$$

Thus, when $s_j = s^f_j,\ \forall s^f_j \in S^f,\ \forall u^g_i \in U,\ \forall G_g \subset G$.

$$b^*_{gj} = \begin{cases} 0, & if\ \eta^*_{ij} = 0,\ \mu^*_{ij} > 0,\ H \neq 0 \\ 1, & if\ \eta^*_{ij} > 0,\ \mu^*_{ij} = 0,\ H = -\eta^*_{ij} \\ \frac{\sum_{u \in U^j} x^*_{ij}}{y_g}, & if\ \eta^*_{ij} = 0,\ \mu^*_{ij} = 0,\ H = 0,\ \sigma^*_{1_{gj}} > 0,\ \sigma^*_{2_{gj}} = 0 \end{cases} \tag{26}$$

Note that $H = \left(\frac{w}{C_{SP_{Max}}}\tau^R_g - \sigma^*_{1_{gj}} + \sigma^*_{2_{gj}}\right)$.

APPENDIX D
PROOF OF POTENTIAL GAME

In [38], the objective functions of the players constitute a potential game with a potential function. We consider the global objective function (potential function) given by Eq. (27) for our problem:

$$\Phi(X,B) := \sum_{j=1}^{m} \left(\frac{w\ C_{SP_j}}{C_{SP_{Max_j}}} + (1-w) T^{resp}_{norm_j} \right) \tag{27}$$

We then assign for each player (home service provider) an objective function that captures the player's marginal contribution to the potential function, as given by Eq. (28).

$$UF_j(X_j, B_j, X_{-j}, B_{-j}) = \sum_{s_i^h \in S^h \setminus s_j^h} \left(\frac{w\ C_{SP_i}}{C_{SP_{Max_i}}} + (1-w) T^{resp}_{norm_i} \right) \tag{28}$$

Note that each player's objective function is only dependent on the actions of other players. Thus, from Eq. (27) and Eq. (28), the potential game of the problem is defined as given by Eq. (29).

$$\Phi_{(X_{-j},B_{-j})}(X_j, B_j) = \sum_{S_j^h \in \mathcal{S}^h} \sum_{S_i^h \in \mathcal{S}^h \setminus S_j^h} \left(\frac{w\ C_{SP_i}}{C_{SP_{Max_i}}} + (1-w) T^{resp}_{norm_i} \right) + \sum_{S_i^h \in \mathcal{S}^h \setminus S_j^h} \left(\frac{w\ C_{SP_i}}{C_{SP_{Max_i}}} + (1-w) T^{resp}_{norm_i} \right) \tag{29}$$

The change in the objective function of player $S_j^h \in \mathcal{S}^h$ by switching from action $(X_j^{'}, B_j^{'})$ to action $(X_j^{''}, B_j^{''})$, provided that all other players collectively play (X_{-j}, B_{-j}), is given by Eq. (30):

$$UF_j(X_j^{''}, B_j^{''}, X_{-j}, B_{-j}) - UF_j(X_j^{'}, B_j^{'}, X_{-j}, B_{-j}) = \Phi_{(X_{-j},B_{-j})}(X_j^{''}, B_j^{''}) - \Phi_{(X_{-j},B_{-j})}(X_j^{'}, B_j^{'}) \tag{30}$$

Consequently, we can deduce that the game is the exact potential game, which always has a Nash equilibrium.