A Survey on Relay-Assisted Computation Offloading Methods in Mobile Edge Computing: Postprint

Authors: Chen Che, Zheng Yifeng, Yang Jingmin, Xie Lingfu, Wenjie Zhang

Date: 2022-04-07T15:01:57+00:00

Abstract

To address the challenges of next-generation networks in coverage, deployment cost, and capacity, Mobile Edge Computing (MEC) often necessitates the assistance of relay nodes to execute computation-intensive and delay-sensitive tasks. This paper first introduces the fundamental architecture of relay-assisted MEC systems, then systematically reviews the latest research methodologies from three perspectives: task offloading, resource allocation, and relay node selection. Furthermore, it discusses and analyzes potential issues and challenges in existing approaches, and proposes viable solutions to serve as references for future research directions.

Full Text

Preamble

Vol. 39 No. 8
Application Research of Computers
ChinaXiv Partner Journal

Survey on Relay-Assisted Computing Methods in Mobile Edge Computing

Chen Che[1],[2], Zheng Yifeng[1],[2], Yang Jingmin[1],[3], Xie Lingfu[4], Zhang Wenjie[1],[2]†

[1]College of Computer Science, Minnan Normal University, Zhangzhou, Fujian 363000, China
[2]Key Laboratory of Data Science & Intelligence Application, Fujian Province University, Zhangzhou, Fujian 363000, China
[3]College of Electronic Engineering, Taipei University of Technology, Taipei

106344, China
[4]College of Electrical Engineering & Computer Science, Ningbo University, Ningbo, Zhejiang 315211, China

Abstract: To address the challenges of next-generation networks in coverage, deployment cost, and capacity, Mobile Edge Computing (MEC) often requires the assistance of relay nodes to complete computation-intensive and latency-sensitive tasks. This paper first introduces the basic architecture of relay-assisted MEC systems, then summarizes the latest research methods for relay-assisted MEC systems from three aspects: task offloading, resource allocation, and relay node selection. Furthermore, it discusses and analyzes potential problems and challenges in existing methods, and proposes some feasible solutions to provide references for future research and development.

Key words: mobile edge computing; relay-assisted; resource allocation; task offloading; relay selection

0 Introduction

In recent years, with the rapid development and popularization of Internet of Things (IoT) and 5G communication technologies, the number of terminal devices has grown exponentially. These application scenarios impose higher requirements on latency and device energy consumption, while also posing tremendous challenges to communication network and computing resource management [1]. Examples include smart healthcare, autonomous driving, and Augmented Reality (AR). Most current mobile terminal devices have small form factors, insufficient computing power, limited resource storage, and battery capacity, making them unable to meet the high computational demands of new applications and seriously affecting device performance and user experience. To address these issues, cloud computing technology concentrates data computing capabilities, storage, and network management in the cloud, primarily involving data centers, backbone IP networks, and cellular core networks [2,3]. It allows devices to upload local applications to the cloud for computation, which not only reduces task execution latency but also decreases energy consumption for local task execution. However, cloud computing has a significant drawback: when the propagation distance between terminal devices and the central cloud network is long, data transmission over the link incurs substantial communication latency, and the probability of transmission interruption increases, making it unable to meet the latency and reliability requirements of sensitive services such as intelligent driving and augmented reality. Consequently, the concept of Mobile Edge Computing (MEC) has been proposed by the industry [4].

Currently, MEC has become one of the key technologies for 5G network development. Unlike traditional cloud computing, MEC deploys servers at network edge nodes closer to terminal devices, localizing services. Its advantages mainly include low latency, low power consumption, and high energy efficiency. While

MEC can improve network performance to some extent, it still faces issues such as limited edge node resources, mismatched communication distances and signal transmission power, and underutilization of idle servers. To fully utilize network resources, cooperative communication is considered an effective solution. Based on cooperative communication scheme design, researchers have proposed relay-assisted MEC systems [5]. Such systems can be divided into three parts: source subsystem, relay subsystem, and destination subsystem, where relays enable indirect communication between source and destination nodes by sharing resources. Unlike traditional MEC systems, relay-assisted MEC systems can not only effectively increase wireless network access capability and expand coverage but also fully utilize idle server resources to improve network energy efficiency and support computation-intensive and latency-sensitive applications, thereby liberating resource-limited mobile devices from these tasks. For example, in VR games, users A and B need to complete specific computing tasks to win the game, and actions taken by either user inevitably affect the final outcome. Therefore, users A and B need to share computation results through a relay-assisted MEC system. To fully utilize idle resources at the edge and improve system coverage and computing power, scholars have conducted in-depth research on relay-assisted MEC systems, focusing on computation task offloading, resource allocation, and relay selection. Computation task offloading and resource allocation mainly address questions of how to offload, when to offload, and how much to offload between user terminals and relays, as well as between relays and destinations [6]. Many researchers have further explored resource scheduling and auction problems in the Internet of Things (IoT). For instance, in [7], a multi-objective distributed scheduling model was proposed and solved using a multi-objective intelligent algorithm based on sine functions, providing new ideas for data processing in IoT. In [8], Cui et al. proposed a subspace clustering post-processing strategy considering sparsity and connectivity, and experiments demonstrated its effectiveness in IoT and ability to improve clustering accuracy. In [9], Zi et al. proposed a lightweight heuristic solution for fast scheduling that can significantly reduce task execution latency in mobile edge networks through task scheduling. In [10], the resource management problem between MEC and devices was modeled as a double auction game process, achieving Nash equilibrium by introducing participants' experience-weighted values. Relay node selection affects performance metrics such as energy consumption, latency, throughput, and transmission interruption probability in cooperative communication systems, making it a noteworthy research problem. Relay cooperation can not only effectively utilize idle resources within the system but also help improve network communication performance. While numerous survey articles exist on MEC basic principles, technological development, and applications, there are relatively few survey articles in the relay-assisted domain. This paper provides a detailed summary of key research issues in relay-assisted MEC systems, including task offloading, resource allocation, and relay node selection, discusses the advantages and disadvantages of existing methods, and prospects several challenges that remain unresolved in current work.

This paper first provides a detailed summary of MEC basic concepts and relay-assisted MEC systems. Second, it summarizes the latest research methods for relay-assisted MEC systems from three aspects: task offloading, resource allocation, and relay selection, discusses the shortcomings and challenges in existing methods, and further proposes some potentially feasible solutions to provide references for future research and development.

1 Relay-Assisted MEC System

This section mainly elaborates on the basic concepts and framework of MEC, then details the relay-assisted MEC system.

1.1 Traditional MEC Framework

MEC is regarded as a small cloud service platform operating at the edge network, which moves computing services originally located in the cloud center to network edge nodes. This not only reduces network operations and communication losses but also helps improve user service quality. Its structure can be divided into three parts: device terminal, edge, and core cloud [11,12], as shown in Figure 1. These three parts are described in detail below.

Device Terminal: Terminal devices (such as mobile phones, laptops, etc.) are deployed in edge computing environments. The edge computing environment can provide users with more interactive communication and response capabilities. With the emergence of massive new applications, a large number of terminal devices in edge computing environments need to provide efficient and real-time computing services for new applications. However, most application computing tasks are large in volume and have low latency requirements, which existing terminal devices cannot meet. Therefore, terminal devices must offload computing tasks to edge servers.

Edge: Edge servers are deployed at the network edge to provide users with close-range, low-latency, low-power, and high-energy-efficiency services, which can further improve network performance. Edge servers can respond to many user service requests, such as data caching, computation task offloading, and data forwarding. Therefore, in edge computing environments, most users migrate tasks to the edge for completion, which not only reduces transmission latency but also significantly improves device performance to meet user needs.

Core Cloud: Core cloud servers can provide sufficient computing power and data storage capabilities for massive data processing. For example, cloud servers can provide services such as massive parallel data processing, machine learning, data mining, and deep learning. However, due to constraints, cloud servers are usually located far away, consuming substantial energy and incurring significant latency in transmission links.

In 5G communication networks, edge devices are deployed more densely, bringing users closer to edge servers and effectively reducing channel loss for task offloading. Since edge nodes (including base stations, laptops, tablets, and mobile phones) have different demands, some devices are idle during each time period while others have overloaded computing tasks. Therefore, MEC systems need to improve resource utilization to meet important requirements in data optimization, agile connectivity, application intelligence, and real-time services. MEC servers typically consist of small network centers deployed by cloud and telecom operators near user ends, connected to the cloud center through gateways. Although their computing capacity is lower than cloud center servers, their advantage lies in being closer to users, helping provide lower latency and more convenient services.

1.2 Relay-Assisted MEC System

Nowadays, traditional MEC has been extensively studied with many relevant achievements. Since MEC moves computing storage capabilities and service capabilities to the network edge, when the number of users increases and computing tasks become heavy, edge nodes may become overloaded and unable to complete large-scale data transmission and computation, preventing substantial resolution of low-latency and high-energy-consumption issues in MEC. Particularly in the 5G and IoT era, mobile application data is constantly updated and generated, urgently requiring massive resources for transmission and processing. On the other hand, 5G networks consist of numerous wireless devices with computing and communication resources, and each device is likely surrounded by other devices with unused or extra resources. Therefore, how to fully utilize these idle device resources is an effective way to address the problems of excessive users, heavy computing tasks, and overloaded edge nodes in the 5G and IoT era. Additionally, some users may be far from edge servers and outside server coverage. To enable remote users to utilize MEC server computing resources, a relay node should be selected to complete computation task transmission. Consequently, relay-assisted task offloading technology has attracted widespread scholarly attention, mainly consisting of three parts [13,14]: source node, relay node, and remote node, as shown in Figure 2. The blue solid line indicates that most users rely on the relay forwarding function of repeaters to distribute tasks to remote nodes for processing. The red dashed line indicates that idle mobile devices can also serve as relay nodes for task forwarding. The yellow dotted line indicates that when communication base stations are busy, they can also act as relay nodes to request assistance from remote base stations. The descriptions for source node, relay node, and remote node are as follows:

Source Node: Composed of terminal devices, responsible for uploading their own computing tasks to relay nodes or remote nodes for processing.

Relay Node: Idle device terminals, repeaters, and edge servers can all serve as relay nodes for data forwarding and computation. Due to limited computing resources, relay nodes cannot provide endless computing services for all devices within their range and need to further offload additional tasks to connected servers for execution.

Remote Node: Collaborates with relay nodes to provide computing services for terminal devices. Because the distance from terminal devices is far and there is no direct connection channel between them, forwarding through relay nodes is required.

Traditional MEC solutions cannot be directly applied in environments with long distances or significant signal interference and require assistance from relay nodes to complete [15]. Compared with traditional MEC, relay-assisted MEC systems can provide cooperative communication and computing services, reduce the load on nearby servers, and improve resource utilization of remote servers. Through cooperative communication and computing via relay nodes, relay-assisted MEC systems can not only expand communication range, reduce signal transmission interference, and decrease the probability of signal interruption but also offload tasks to remote idle servers when nearby resources are limited, effectively improving energy efficiency in MEC environments.

The simplest relay-assisted MEC system contains three nodes: one user node, one relay node, and one remote node. In this scenario, only offloading problems such as power allocation and computation task volume need to be considered. However, in practical scenarios, MEC systems typically have multiple users and multiple relay nodes, which poses significant challenges for relay selection and task offloading. Research shows that in a system with multiple relay nodes, selecting either a single best relay node or multiple relay nodes for data transmission can achieve diversity gain and improve system performance [16]. Based on obtaining channel state information and the computing and forwarding capabilities of candidate relay nodes, traditional relay selection mainly includes partial relay selection and opportunistic relay selection. The former obtains local channel information from source nodes to relay nodes, while the latter utilizes global channel information from source nodes to relay nodes and from relay nodes to remote nodes for relay selection. Compared with the latter, the former's solution results cannot achieve global optimality but have lower complexity. Considering scenarios with a large number of users in MEC environments, serious signal transmission interference can easily occur in Device-to-Device (D2D) communications due to long distances between users [17]. Additionally, due to limited resources and diverse user demands, competition exists among users when multiple users select the same relay node. Therefore, how to reasonably allocate resources and execute task offloading based on relay node resources, MEC network environment, terminal offloading requests, and resource conditions to improve resource utilization and meet the needs of more users will be a major challenge for relay-assisted MEC systems.

The following sections will elaborate in detail on research methods for relay-assisted MEC systems focusing on three aspects: resource allocation, task offloading, and relay selection.

2 Relay-Assisted Task Offloading

Task offloading, as one of the key technologies in MEC, transmits computing tasks to edge service nodes for execution, thereby alleviating computation and extending device battery life. In MEC, task offloading is mainly divided into full offloading and partial offloading [5]. Through full offloading, computing tasks on mobile devices are entirely offloaded to edge servers for computation or executed directly on local devices. Through partial offloading, computing tasks can be divided into multiple different parts for computation on edge servers and local devices respectively. In [18], a low-complexity algorithm was proposed to solve the full offloading problem under delay constraints. In [19], Sun et al. aimed to maximize the sum of user computation efficiency with weighted factors, using iterative and gradient descent methods to trade off local computation and task offloading. Since full offloading tasks are indivisible, partial offloading is more suitable for parallel processing tasks. In [20], a semidefinite relaxation-based algorithm was used to solve partial offloading decision problems, effectively reducing offloading energy consumption and delay. In [21], a deep reinforcement learning-based adaptive computation offloading method was proposed, which can effectively avoid the problem of excessively complex action spaces in time-varying environments and effectively learn optimal strategies. In relay-assisted MEC systems, task offloading is more flexible. Each task can be processed locally, offloaded to relay nodes for execution, or forwarded through relay nodes to remote nodes for execution. As shown in Figure 3, subfigure (a) represents users executing tasks locally, subfigure (b) represents users offloading small-scale tasks to nearby relay nodes for processing, and subfigure (c) represents users offloading large-scale tasks through relay nodes to remote servers for execution. Different offloading decisions produce different transmission delays and energy losses. For example, local task processing has no transmission delay but may result in higher task execution time and energy consumption, while offloading tasks to nearby edge servers or remote servers can reduce task execution delay and energy consumption to a certain extent but inevitably leads to additional transmission delay. Therefore, how to make reasonable offloading decisions and offload computing tasks to edge nodes has become one of the important research directions in MEC, with significant research importance in shortening task delay and improving energy efficiency.

After introducing relay nodes into MEC systems, users can utilize relay nodes to process computing tasks, making user task offloading decisions more attractive and attracting many scholars to study. In existing relay-assisted task offloading methods, the main optimization metrics are minimizing delay or energy consumption. Therefore, using performance as a classification criterion, user task offloading methods can be divided into three categories: delay-minimization offloading methods, energy-minimization offloading methods, and multi-objective optimization offloading methods, as shown in Table 1.

Delay-Minimization Offloading Methods: Many large-scale applications today have ultra-low latency requirements, such as video surveillance and smart transportation. Therefore, minimizing task execution delay has always been a primary concern in MEC systems. For example, in delay-minimization task offloading methods, Meng et al. transformed the optimization problem of minimizing average slowdown and average timeout of tasks in buffer queues into a learning problem, proposing a Deep Reinforcement Learning (DRL)-based algorithm and designing a reward function to guide the algorithm to learn offloading strategies directly from the environment [22]. Xing et al. proposed a task offloading method based on Time Division Multiple Access (TDMA) communication protocol in multi-user cooperative MEC systems, reducing local user computation delay while optimizing task allocation, time allocation, and power allocation. Since the joint task allocation and wireless resource allocation problem is a Mixed-Integer Non-Linear Programming (MINLP) problem, it must first be relaxed into a convex problem to obtain an effective suboptimal solution [23]. Additionally, [24] offloaded computing tasks to multiple destinations with computing capabilities and proposed multiple destination selection criteria to maximize the computing capacity of relay nodes, channel gain of relay links, and channel gain of direct links. To better evaluate system performance related to delay, an outage probability based on transmission and computation time was proposed. Experiments showed that relay-assisted MEC can significantly alleviate the impact of increasing computing tasks and reduce delay.

Energy-Minimization Offloading Methods: Energy consumption has always been a key focus in MEC systems, with long-term significance for extending device operation time and battery life and protecting the environment. Generally, adjusting computing task offloading schemes can reduce the energy consumption required to complete tasks to a certain extent. For example, Peng et al. proposed an adaptive offloading compression energy-saving mechanism (ESAOC) to solve the problems of excessive energy consumption and communication overhead in cloud-enhanced fiber-wireless networks [25]. First, the offloading compression ratio of services was dynamically adjusted by combining the average arrival rate of offloading data with different priorities and the compression delay of each node. Then, a queuing model was established to analyze the queuing delay of offloading services in MEC servers, and wireless side relay nodes were cooperatively scheduled for coordinated sleep scheduling. Li et al. studied the total energy consumption minimization problem in multi-relay-assisted MEC systems under TDMA mode, Decode-and-Forward (DF), and Amplify-and-Forward (AF) scenarios in FDMA mode, solving it through bi-level optimization [26]. Additionally, [27] considered a computing architecture with multiple sources, multiple relays, and a single edge server, dividing the minimization of maximum energy consumption problem in OFDMA wireless networks into two subproblems, and proposing an Optimal Total energy Consumption Algorithm (OTCA) based on bipartite graph matching and an Optimal Energy Consumption Allocation Algorithm (OECAA) to solve them respectively. [28] proposed an offloading method based on a harvest-then-offload protocol, converting the problem of minimizing the total transmission energy of access points while satisfying computing task constraints into a minimization of maximum energy consumption problem, and obtaining optimal offloading decisions and optimal minimum energy transmission power through solving.

Multi-Objective Optimization Offloading Methods: In task offloading methods that trade off multiple objectives, most scholars mainly study joint optimization of delay and energy consumption. Additionally, researchers choose different offloading schemes based on different Quality of Service (QoS) indicators and performance requirements. For example, Fan et al. considered a multi-server MEC system, proposed a task scheduling optimization method based on minimizing delay and energy consumption, and used balance factors to flexibly adjust the optimization deviation between delay and energy consumption [29]. Wen et al. studied a Multi-Input Multi-Output (MIMO) full-duplex (FD) relay-assisted SWIPT-MEC system to reduce system energy consumption, proposing an energy efficiency problem to minimize energy consumption within a time period while ensuring delay constraints and energy consumption constraints [30]. Ranji et al. proposed an Energy-Efficient and Delay-Aware Offloading Scheme (EEDOS) based on D2D collaboration in MEC. First, offloading requests were classified according to the deadline and energy constraints of requesting terminal devices, then suitable offloading destinations were found through maximum matching of a maximum-cost graph algorithm [31]. Cao et al. proposed a joint computation-communication cooperation method under different offloading methods to jointly optimize the computation and communication resource allocation of users and relay nodes (i.e., offloading time and transmission power allocation and CPU frequency) while satisfying user computation delay constraints, minimizing their total energy consumption [32]. Additionally, Liao et al. proposed a wireless relay-supported task offloading mechanism based on a medical monitoring framework, focusing on performance evaluation under certain link quality conditions, and then selecting different computation offloading schemes according to different QoS indicator requirements [33]. Dong et al. constructed an intelligent mobile edge computing (NOMA-MEC) communication system based on cooperative non-orthogonal multiple access, proposing performance analysis of offloading for intelligent mobile edge computing systems based on NOMA, and evaluating the performance of cooperative NOMA-MEC systems through expressions of offloading outage probability for a pair of users [34].

Delay-minimization offloading methods have great advantages in reducing delay, but in some scenarios with low delay requirements, they may cause excessive network energy consumption overhead due to over-pursuing delay. Similarly, energy-minimization offloading methods can effectively reduce energy consumption in the network, but in some application scenarios with high delay requirements, they cannot meet low-latency computation needs, leading to task dropping. Multi-objective optimization offloading methods comprehensively consider the problems existing in single-objective optimization. However, nowadays, most literature can only obtain approximate solutions to problems, which still have a large gap from the optimal solutions obtained by exhaustive search methods.

3 Relay-Assisted Resource Allocation

Resource allocation problems in MEC mainly include: 1) **Computing resource allocation:** Effectively allocating computing resources according to user task requirements to minimize energy consumption and task execution delay; 2) **Communication resource allocation:** Allocating communication resources according to the network environment to optimize system transmission efficiency and reduce communication interference among users. In MEC, reasonable resource allocation can fully utilize idle resources and is therefore crucial. In [35], Li et al. proposed a reinforcement learning-based optimization framework to solve resource allocation problems in MEC, significantly reducing user delay and energy consumption. In [36], an adaptive computing resource allocation scheme was proposed, which can balance network resources in different MEC environments and outperform traditional algorithms. In relay-assisted MEC systems, using nearby edge nodes as relays to assist source nodes in forwarding tasks to destination nodes can fully utilize idle resources of surrounding nodes and improve overall network performance. Additionally, relay-assisted resource allocation problems need to consider not only user task requirements and MEC server resources but also available resources of relay nodes and distances between relay nodes and source/destination nodes, making them more difficult than traditional MEC resource allocation problems. In relay-assisted MEC systems, users can achieve higher capacity with lower resources by selecting appropriate relay nodes for transmission. Relay-assisted resource allocation strategies mainly depend on the amount of user-requested tasks and their own available resources. The basic allocation process is shown in Figure 4. First, user terminals completely offload applications and deliver them to local schedulers within MEC. The scheduler checks for computing nodes with sufficient resources. If a relay node capable of processing the application exists, resources are allocated on that node to process the application and send the results to the user. If no suitable computing node exists, the application is forwarded with the help of relay nodes to remote servers. Currently, relay node-based resource allocation methods mainly use delay and energy consumption as performance indicators. Similarly, server resource allocation methods can be divided into three categories: delay-minimization resource allocation methods, energy-minimization resource allocation methods, and multi-objective optimization resource allocation methods, as shown in Table 2.

Delay-Minimization Allocation Methods: Zhang et al. studied resource scheduling problems in UAV-assisted MEC systems, including UAV flight trajectory, association between UAV hovering positions and UEs, and task scheduling, aiming to minimize task completion time, reduce system delay, and improve user experience [37]. Qin et al. proposed a communication resource scheduling and multi-user computation uploading problem based on minimizing task processing delay, quantified user-available computing resources, derived a system utility maximization function, further deduced subcarrier allocation and power allocation criteria in multi-user MEC systems, and then proposed specific resource allocation algorithms [38].

Energy-Minimization Allocation Methods: To effectively reduce total system energy consumption and meet the delay requirements of smart devices, [39] jointly optimized D2D-based relay selection and resource allocation strategies in MEC systems and proposed a two-stage optimization algorithm to solve this mixed-integer non-convex optimization problem. First, convex optimization techniques were used to transform the original problem into a convex problem to obtain the optimal relay selection strategy. Then, the Lagrangian method was used to transform the original problem into a resource allocation problem. Under constraints of computation, communication, and delay, [40] proposed joint optimization of relay selection strategy and resource allocation strategy to optimize mobile device energy consumption. First, convex optimization techniques were used to transform the problem into a mixed-integer non-convex optimization problem, then the optimal resource allocation strategy was obtained through Lagrange multiplier method and relay selection strategy. Aiming to minimize energy consumption, [40] proposed a D2D-assisted MEC system for long-distance computation task offloading in MEC. Under constraints of computation, delay, and communication, the optimization problem of minimizing smart mobile device energy consumption was achieved through joint optimization of resource allocation strategy and relay selection strategy, realizing energy-efficient relay routing selection and resource allocation strategies.

Multi-Objective Optimization Allocation Methods: Chen et al. [41] proposed a resource allocation method based on hybrid relay forwarding protocols, modeling the minimization problem of execution delay and network energy consumption as a highly coupled constrained non-differentiable non-convex optimization problem, and proposed a lightweight algorithm based on inexact Block Coordinate Descent (BCD) to solve it. Tan et al. [42] jointly considered user association, power control, and resource allocation (including spectrum, cache, and computation) in MEC networks with high data rate services and computation-sensitive services, studied virtual resource allocation for heterogeneous services in full-duplex small cell networks (SCN), and obtained optimal solutions through variable relaxation and Alternating Direction Method of Multipliers (ADMM) algorithm. Additionally, for relay-assisted systems, Amplify-and-Forward (AF) and Decode-and-Forward (DF) at relay nodes may also have significant impacts on network performance and user Quality of Experience (QoE). Based on the above considerations, [43] proposed a new hybrid relay (HR) architecture for relay-assisted computation offloading (RACO) and designed an effective resource allocation method for relay-assisted task offloading, thereby reducing task execution delay and energy consumption. Under practical constraints of available computing and communication resources, this method minimizes the weighted sum of execution delay and energy consumption in the RACO system by jointly optimizing the computation offloading ratio, processor clock rate, bandwidth allocation between AF and DF schemes, and transmission power levels of users and mobile edge relay servers (MERS). [44] studied the joint problem of cooperative computation task offloading schemes and resource allocation in MEC, reducing delay under constraints of transmission power, energy consumption, and CPU cycle frequency, and solving this non-convex mixed-integer problem step by step using conceptual optimization methods, ShengJin Formula method, and monotonic optimization methods.

Similar to offloading methods, delay-minimization resource allocation methods can reasonably allocate computing resources to achieve the goal of minimizing delay, but these methods ignore energy consumption issues at different computing nodes. Energy-minimization methods have difficulty meeting delay requirements. Therefore, how to trade off between energy consumption and delay remains challenging.

4 Relay Selection

Relay selection refers to choosing an optimal relay node from all candidate relay node sets for data forwarding and task offloading, as shown in Figure 5. Cooperative communication and computation within MEC systems need to rely on relay node assistance. Due to limitations in edge node computing capacity and channel conditions, they may be unable to provide computing services for users' large tasks. Therefore, selecting the optimal relay node can effectively improve MEC system performance and expand MEC system coverage. This subsection provides a detailed overview of relay node selection methods from the perspectives of communication assistance and computation assistance.

Communication-Assisted Relay Selection Methods: When MEC resources are limited, communication distance and signal transmission power intensity are mismatched, and idle servers are not fully utilized, relay nodes are needed for assisted communication. Communication assistance can effectively save energy, improve network throughput, and enhance network coverage. For example, [45] used relay node cooperative MIMO (CMIMO) to transmit data from nodes to base stations, effectively solving the problem of deep channel fading between source nodes and destinations. Using clustering and data aggregation methods, cluster heads were selected in each cluster as relay nodes to aggregate and forward data. By constructing energy consumption system models for inter-cluster and intra-cluster communication and defining some cooperative communication schemes, data was transmitted from relay nodes to destinations through another cooperative node in fading channels to reduce energy consumption in the system. [46] considered relay selection problems equipped with battery energy storage, studying relay selection strategies corresponding to different complexities under several different Channel State Information (CSI) requirements, including random relay selection strategy, distance-based best relay selection strategy, and relay selection strategy considering battery performance. Analytical results for the outage probability performance of the proposed schemes and simplified asymptotic expressions for high Signal-to-Noise Ratio (SNR) regimes were derived. In [47], Jing et al. proposed a relay selection method allowing multiple relay offloading. Based on a relay ranking function, multiple SNR suboptimal (and thus error rate suboptimal) relay selection schemes were proposed to achieve full diversity and low error rates. In [48], Zhao et al. jointly considered channel conditions between source nodes and relay nodes and between relay nodes and destination nodes to reduce bit error rate, proposing schemes for cases where the number of relay nodes is greater than and less than a threshold. When the number of relay nodes is less than the given threshold, the scheme focuses on the accuracy of node selection. When the number of relay nodes exceeds the threshold, it focuses on the effectiveness and complexity of node selection. Simulation experiments show that the proposed relay node selection scheme can reduce the bit error rate while maintaining unchanged outage probability compared with existing schemes.

Computation-Assisted (D2D) Relay Selection Methods: D2D (Device-to-Device) refers to direct communication between two devices in close proximity without requiring a base station. This can effectively reduce delay and energy consumption. Comprehensive consideration of computation assistance and D2D technology can not only improve resource utilization but also help enhance energy efficiency and communication capacity. In [39], Li et al. minimized the total energy consumption of D2D MEC systems by jointly optimizing wireless communication, relay selection, and computing resource allocation in MEC systems. The system model allowed each user to select a relay node from a corresponding relay set to offload data and simultaneously determined the offloading ratio based on different relay node selections. Considering constraints of computation, communication, delay, and selection, the problem could be described as a mixed-integer non-convex optimization problem and is an NP-hard problem. Since enumerating all possible relay selection strategies has high time complexity, to solve this problem, the discrete variables of relay selection could be relaxed to construct new constraints. Then, reformulation-linearization techniques were used to eliminate all quadratic terms of selection decisions. Through these two steps, a standard convex problem was obtained, and the optimal solution was obtained by solving this convex problem. Simulation results show that the proposed algorithm not only reduces the time to obtain optimal solutions but also helps improve energy efficiency. In [49], Rahman et al. selected the best D2D relay through an Adaptive Neuro-Fuzzy Inference System (ANFIS) architecture to forward D2D source information to the intended D2D destination, proposing an adaptive network ANFIS architecture using supervised learning. The Signal-to-Noise Ratio (SNR) of D2D communication and the SNR of relay transmission to remote servers served as two input variables entering a five-layer ANFIS architecture. The first layer nodes used fuzzification to generate membership grades; the second layer nodes generated firing strengths; the third layer calculated qualified subsequent membership functions (MF) based on firing strengths; the fourth layer generated the overall output membership function; and the fifth layer obtained clear output values through defuzzification. The proposed ANFIS architecture constructed a mapping model between D2D communication relay selection parameters and factors. The neuro-fuzzy-based algorithm selected relay nodes with maximum D2D communication energy efficiency, whose energy efficiency was close to the exhaustive search-based best relay selection scheme. In [50], Omran et al. considered the impact of relay selection on load balancing, formulating the problem as a mixed-integer combinatorial optimization problem, which is NP-hard. Therefore, a joint user relay selection scheme with load balancing for multi-layer heterogeneous networks was proposed, using the Kuhn-Munkres (K-M) method to give the final selection, thereby achieving dynamic load balancing based on D2D communication. Simulation results show that this scheme can significantly increase the number of admitted offloading users under load balancing conditions.

Communication + Computation-Assisted Relay Selection Methods: Relay-assisted computation methods can not only alleviate load pressure within MEC systems but also help improve network coverage. Therefore, most research on relay-assisted MEC systems adopts a combination of communication and computation assistance. For example, in [51], Chen et al. proposed a Sequential Relay-remote Selection and Offloading (SRSO) strategy to minimize energy consumption by specifying when to stop node sequence discovery and execute computation offloading. This paper distributed computing tasks to local, relay, and remote parts to improve parallel execution capability of task computing. In the two processes of offloading computing tasks from local to Relay Node (RN) and from RN to Remote Server (RS), the system made optimal decisions based on the system state (i.e., node computing capacity and channel conditions) at each stage of the current process to determine the offloading ratio. Taking the local-to-RN offloading process as an example, the system made optimal decisions based on the current state (i.e., relay node computing capacity and channel conditions). This could be expressed through formulas as follows:

[Formula placeholders appear to be incomplete or corrupted in the original text]

If relay nodes have task buffer queues, the queue length can also be considered in relay selection decisions. This scheme can be regarded as Weighted Transmission energy and Buffer queue Minimization (WTBM), as shown below:

[Additional formula placeholders appear incomplete]

In [52], Liang et al. introduced relay selection schemes considering only communication capability (Communication-Only Relay Selection, CORS) and only computing capability (Computing-Only Selection, CPORS). When only communication is considered, the relay selection scheme is based on transmission rates between users and relay nodes and between relay nodes and destinations. When only computing capability is considered, the relay selection scheme is based on the constant CPU cycle frequency of relay nodes, where higher frequency indicates stronger computing capability. [52] further proposed the Latency-Best Relay Selection (LBRS) scheme, which considers not only communication capability but also the computing capability of relay nodes. Therefore, the relay node selected by the LBRS scheme minimizes the sum of transmission delay from user to relay node, computation delay at relay node, and transmission delay from relay node to destination.

In [53], Li Taoshen et al. proposed a Simultaneous Wireless Information and Power Transfer (SWIPT) relay selection scheme based on power allocation cooperation to solve the nonlinear mixed-integer programming problem constituted by constraints such as channel quality and relay transmission power. Compared with traditional relay cooperation schemes, the cache configuration proposed in this paper is feasible and effective, and the scheme has higher throughput, as shown in Table 3.

5 Problems and Challenges

Although relay-assisted MEC systems have begun to be widely studied and scholars have achieved some research results, the following problems still lack better solutions:

(1) User Mobility and Task Randomness: High user mobility is one of the characteristics of MEC systems. When users move, relay-assisted nodes need to be reselected. Additionally, users do not always have tasks to compute, and tasks arrive randomly with random sizes. Therefore, the number and size of tasks to be processed in MEC systems are variable, which directly affects offloading decisions, resource allocation, and relay selection. Markov approximation is an effective technique for analyzing dynamic models. By designing reversible and convergent Markov chains to analyze user dynamic behavior, user mobility problems can be mapped to state transition problems of Markov chains. Additionally, dynamic distributed algorithms can be considered by dividing stages according to task arrivals.

(2) Relay Path Optimization: When users are far from edge servers, relying on a single relay node may be unable to complete task offloading. How to find an optimal path through multiple relay nodes is a difficult problem. Relay path selection needs to comprehensively consider factors such as channel capacity, path loss, distance, and computing capacity of relay nodes, calculating the energy consumption and delay of each path between users and relay nodes, between relay nodes, and between relay nodes and MEC servers, to select the next forwarding node for tasks and find a path with minimum energy consumption or delay from users to remote servers through multiple relay nodes.

(3) Relay Incentive Schemes: With the application of 5G and IoT, numerous mobile vehicles and drones equipped with edge computing servers exist in the network, which can provide real-time computing services for users. However, these network devices are selfish. When providing relay services, devices need to consume resources such as battery and computing capacity. Additionally, device resource sharing poses risks of exposing location and other private information. Therefore, without satisfactory rewards to compensate for resource consumption and potential privacy breaches, these network devices generally will not actively provide relay services or share resources. Consequently, how to design incentive schemes to attract mobile vehicles, drones, and other wireless devices to become relay nodes and share their idle resources is an effective way to improve energy efficiency and system computing capacity. Additionally, how to consider relay node characteristics (e.g., drone flight paths, drone distances, mobile vehicle speed and direction) when performing task offloading and resource allocation is also a major challenge. Contract theory is an effective method for designing incentive mechanisms in resource markets. Its key idea is to formulate appropriate contract terms and use monetary incentives to attract network devices to provide relay services and share resources.

(4) 5G Networks: 5G networks contain a large number of wireless devices with specific computing and communication resources. Due to the burstiness of wireless communication, each wireless device is likely surrounded by idle devices with unused or extra resources. Therefore, how to combine relay-assisted MEC systems with these idle device resources to improve 5G network computing capacity and resource utilization is an urgent problem to be solved. Additionally, MEC systems can support many types of applications with different requirements for energy consumption, delay, and resources. For example, in health monitoring, since terminal devices are relatively fixed, service demands are long-term, while for some highly mobile devices, service demands are short-term. This is consistent with the 5G network service model that meets more users' personalized needs. Integrating relay-assisted MEC with 5G network slicing technology and introducing Virtual Machine (VM) mechanisms to achieve combined allocation of computing and communication resources can optimize system resource utilization while maintaining required QoE levels. According to relay node resource conditions and user demands, after introducing VM mechanisms, 5G network resource allocation and relay selection problems include the following decision processes: which resources to allocate to users, how much of each resource to allocate, and which relay node to select. Additionally, to meet different users' dynamic service demands during network slicing operation, an adaptive VM algorithm should be designed to transform relay nodes and allocate resources on demand.

6 Conclusion

This paper first introduced the basic concepts and reference architecture of MEC and the basic architecture of relay-assisted MEC systems. Then, it summarized existing methods for relay-assisted MEC systems from three aspects: task offloading, resource allocation, and relay node selection. Finally, it further analyzed several unresolved problems in existing methods and briefly 展望 ed corresponding solution directions.

References

[1] IMT-2020 (5G) Promotion Group. 5G Vision and Demand White Paper V1.0 [EB/OL]. (2014) [2021].

[2] Zhang Qi, Lu Cheng, Boutaba R. Cloud computing: state-of-the-art and research challenges [J]. Journal of Internet Services and Applications, 2010, 1

(1): 7-18.

[3] Armbrust M, Fox A, Griffith R, et al. Above the clouds: A berkeley view of cloud computing [R]. Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, 2009.

[4] Tian Hui, Fan Shaoshuai, Lyu Xinchen, Zhao Pengtao, He Shuo. Mobile Edge Computing for 5G Requirements [J]. Journal of Beijing University of Posts and Telecommunications, 2017, 40 (2): 1-10.

[5] Mao Yuyi, You Changsheng, Zhang Jun, et al. A survey on mobile edge computing: The communication perspective [J]. IEEE Communications Surveys & Tutorials, 2017, 19 (4): 2322-2358.

[6] Xie Renchao, Lian Xiaofei, Jia Qingmin, et al. Overview of mobile edge computing offloading technology [J]. Journal on Communications, 2018, 39 (11): 142-159.

[7] Cai Xingjuan, Geng Shaojin, Wu Di, et al. A multicloud-model-based many-objective intelligent algorithm for efficient task scheduling in internet of things [J]. IEEE Internet of Things Journal, 2020, 8 (12).

[8] Cui Zhihua, Jing Xuechun, Zhao Peng, et al. A new subspace clustering strategy for AI-based data analysis in IoT system [J]. IEEE Internet of Things Journal, 2021, 8 (16): 12540-12549.

[9] Wang Zi, Zhao Zhiwei, Min Geyong, et al. User mobility aware task assignment for mobile edge computing [J]. Future Generation Computer Systems, 2018, 85 (AUG.): 1-8.

[10] Li Quanyi, Yao Haipeng, Mai Tianle, et al. Reinforcement-learning-and belief-learning-based double auction mechanism for edge computing resource allocation [J]. IEEE Internet of Things Journal, 2019, 7 (7).

[11] Ke Hongchang, Wang Jian, Deng Lingyue, et al. Deep reinforcement learning-based adaptive computation offloading for MEC heterogeneous vehicular networks [J]. IEEE Transactions on Vehicular Technology, 2020, 69 (7): 7916-7929.

[12] Xia Shichao, Yao Zhixiu, Xian Yongju, et al. Distributed heterogeneous task offloading algorithm in mobile edge computing [J]. Journal of Electronics and Information Technology, 2020, 42 (12): 68-75.

[13] Naeem M, Anpalagan A, Jaseemuddin M, et al. Resource allocation techniques in cooperative cognitive radio networks [J]. IEEE Communications Surveys & Tutorials, 2013, 16 (2): 729-744.

[14] Tang Xuanxuan, Yang Wendong, Cai Yueming, et al. Overview of buffer-assisted relay selection schemes in cooperative communication [J]. Military Communication Technology, 2017, 038 (001): 35-40.

[15] Zhao Zichao, Zhao Rui, Xia Junjuan, et al. A Novel Framework of Three-Hierarchical Offloading Optimization for MEC in Industrial IoT Networks [J]. IEEE Transactions on Industrial Informatics, 2020, 16 (8).

[16] Asshad M, Khan S A, Kavak A, et al. Cooperative communications using relay nodes for next-generation wireless networks with optimal selection techniques: A review [J]. IEEJ Transactions on Electrical and Electronic Engineering, 2019, 14 (5): 658-669.

[17] Li Peng, Guo Song. Literature survey on cooperative device-to-device communication [J]. Cooperative Device-to-Device Communication in Cognitive Radio Cellular Networks, 2014: 7-12.

[18] Deng Maofei, Tian Hui, Fan Bo. Fine-granularity based application offloading policy in cloud-enhanced small cell networks [C]// 2016 IEEE International Conference on Communications Workshops (ICC). IEEE, 2016: 638-643.

[19] Sun Haijian, Zhou Fuhui, Hu R Q. Joint offloading and computation energy efficiency maximization in a mobile edge computing system [J]. IEEE Transactions on Vehicular Technology, 2019, 68 (3): 3052-3056.

[20] Ding Changfeng, Wang Junbo, Cheng Ming, et al. Joint beamforming and computation offloading for multi-user mobile-edge computing [C]// 2019 IEEE Global Communications Conference (GLOBECOM). IEEE, 2019: 1-6.

[21] Ke Hongchang, Wang Jian, Deng Lingyue, et al. Deep reinforcement learning-based adaptive computation offloading for MEC heterogeneous vehicular networks [J]. IEEE Transactions on Vehicular Technology, 2020, 69 (7): 7916-7929.

[22] Meng Hao, Chao Daichong, Guo Qianying, et al. Delay-sensitive task scheduling with deep reinforcement learning in mobile-edge computing systems [C]// Journal of Physics: Conference Series. IOP Publishing, 2019, 1229 (1): 012059.

[23] Xing Hong, Liu Liang, Xu Jie, et al. Joint task assignment and wireless resource allocation for cooperative mobile-edge computing [C]// 2018 IEEE International Conference on Communications (ICC). IEEE, 2018: 1-6.

[24] Xia Junjuan, Li Chao, La Xiazhi, et al. Cache-aided mobile edge computing for B5G wireless communication networks [J]. EURASIP Journal on Wireless Communications and Networking, 2020, 2020 (1): 1-10.

[25] Peng Haiying, Wang Zedong, Wu Dapeng. Cloud-enhanced FiWi network energy-saving mechanism with offload compression incentives [J]. Journal of Electronics & Information Technology, 2020, 42 (7): 1726-1733.

[26] Li Xiang, Fan Rongfei, Hu Han, et al. Energy-efficient Resource Allocation for Mobile Edge Computing with Multiple Relays [J]. IEEE Internet of Things Journal, (2021-09-13). http://doi: 10.1109/JIOT.2021.3125953.

[27] Yao Mianyang, Chen Long, Liu Tonglai, et al. Energy efficient cooperative edge computing with multi-source multi-relay devices [C]// 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2019: 865-870.

[28] Hu Xiaoyan, Wong K K, Yang Kun. Wireless Powered Cooperation-Assisted Mobile Edge Computing [J]. IEEE Transactions on Wireless Communications, 2018, 17 (4): 2375-2388.

[29] Fan Wenhao, Liu Yyuanan, Tang Bihua, et al. Computation Offloading Based on Cooperations of Mobile Edge Computing-Enabled Base Stations [J]. IEEE Access, 2018, 6: 22622-22633.

[30] Wen Zhigang, Yang Kaixi, Liu Xiaoqing, et al. Joint offloading and computing design in wireless powered mobile-edge computing systems with full-duplex relaying [J]. IEEE Access, 2018, 6: 72786-72795.

[31] Ranji R, Mansoor A M, Sani A A. EEDOS: An energy-efficient and delay-aware offloading scheme based on device to device collaboration in mobile edge computing [J]. Telecommunication Systems, 2020, 73 (2).

[32] Cao Xiaowen, Wang Feng, Xu Jie, et al. Joint computation and communication cooperation for energy-efficient mobile edge computing [J]. IEEE Internet of Things Journal, 2018, 6 (3): 4188-4200.

[33] Liao Yangzhe, Yu Quan, Han Yi, et al. Relay-enabled task offloading management for wireless body area networks [J]. Applied Sciences, 2018, 8 (8): 1409.

[34] Dong Xiequn, Li Xuehua, Yue Xinwei, et al. Performance Analysis of Cooperative NOMA Based Intelligent Mobile Edge Computing System [J]. China Communications, 2020, 17 (8): 45-57.

[35] Li Ji, Gao Hui, Lyu Tiejun, et al. Deep reinforcement learning based computation offloading and resource allocation for MEC [C]// 2018 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2018: 1-6.

[36] Wang Jiadai, Zhao Lei, Liu Jiajia, et al. Smart resource allocation for mobile edge computing: A deep reinforcement learning approach [J]. IEEE Transactions on Emerging Topics in Computing, 2019, 9 (3): 1529-1541.

[37] Zhang Bingxin. Research on UAV-based Mobile Edge Computing Resource Scheduling Mechanism [D]. China University of Mining and Technology, 2020.

[38] Qin Min. Computing upload and resource scheduling method in power-constrained edge computing system [D]. University of Science and Technology of China, 2019.

[39] Li Yang, Xu Gaochao, Yang K, et al. Energy Efficient Relay Selection and Resource Allocation in D2D-Enabled Mobile Edge Computing [J]. IEEE Transactions on Vehicular Technology, 2020, 69 (12): 15800-15814.

[40] Li Yang. Research on Several Issues of Energy-saving and High-efficiency Resource Joint Optimization in Mobile Edge Computing [D]. Jilin University, 2020.

[41] Chen Xihan, Shi Qingjiang, Cai Yunlong, et al. Joint Cooperative Computation and Interactive Communication for Relay-Assisted Mobile Edge Computing [C]// 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE, Chicago, IL, USA, 27-30 Aug. 2018.

[42] Tan Zhiyuan, Yu F R, Li Xi, et al. Virtual resource allocation for heterogeneous services in full duplex-enabled SCNs with mobile edge computing and caching [J]. IEEE Transactions on Vehicular Technology, 2017, 67 (2): 1794-1808.

[43] Chen Xihan, Cai Yunlong, Shi Qingjiang, et al. Efficient Resource Allocation for Relay-Assisted Computation Offloading in Mobile-Edge Computing [J]. IEEE Internet of Things Journal, 2020, 7 (3): 2452-2468.

[44] Kuang Zhufang, Ma Zhihao, Li Zhe, et al. Cooperative computation offloading and resource allocation for delay minimization in mobile edge computing [J]. Journal of Systems Architecture, 2021, 118: 102167.

[45] Shikha, Dayal P. Energy efficient different cooperative communication schemes in wireless sensor network: A survey [C]// IEEE India Conference. IEEE, 2015.

[46] Krikidis I. Relay selection in wireless powered cooperative networks with energy storage [J]. IEEE Journal on Selected Areas in Communications, 2015, 33 (12): 2596-2610.

[47] Jing Yindi, Jafarkhani H. Single and multiple relay selection schemes and their achievable diversity orders [J]. IEEE Transactions on Wireless Communications, 2009, 8 (3): 1414-1423.

[48] Zhao Yuli, Guo Li, Zhu Zhiliang, Yu Hai. Design of a relay node selection scheme in cooperative communication [J]. Computer Applications, 2015, 35 (01): 1-4.

[49] Rahman M, Lee Y D, Koo I. Energy-Efficient Power Allocation and Relay Selection Schemes for Relay-Assisted D2D Communications in 5G Wireless Networks [J]. Sensors, 2018, 18 (9).

[50] Omran A, Sboui L, Rong Bo, et al. Joint Relay Selection and Load Balancing using D2D Communications for 5G HetNet MEC [C]// 2019 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2019.

[51] Chen Che, Guo Rongzong, Zhang Wenjie, et al. Optimal sequential relay-remote selection and computation offloading in mobile edge computing [J]. The Journal of Supercomputing, 2022, 78 (1): 1093-1116.

[52] Liang Jie, Chen Zhiyong, Li Cheng, et al. Delay Outage Probability of Multi-relay Selection for Mobile Relay Edge Computing System [C]// 2019 IEEE/CIC

International Conference on Communications in China (ICCC). IEEE, 2019.

[53] Shi Anni, Li Taoshen, Wang Zhe, et al. Relay selection strategy of full-duplex wireless energy-carrying communication system based on buffer assist [J]. Computer Applications, 2021, 41 (06): 1539-1545.

Note: Figure translations are in progress. See original paper for figures.