Task Offloading with Multi-Tier Computing Resources in Next Generation Wireless Networks

Kunlun Wang, *Member, IEEE*, Jiong Jin, *Member, IEEE*, Yang Yang, *Fellow, IEEE*, Tao Zhang, *Fellow, IEEE*, Arumugam Nallanathan, *Fellow, IEEE*, Chintha Tellambura, *Fellow, IEEE*, and Bijan Jabbari *Fellow, IEEE*

***Abstract*—With the development of next-generation wireless networks, the Internet of Things (IoT) is evolving towards the intelligent IoT (iIoT), where intelligent applications usually have stringent delay and jitter requirements. In order to provide low-latency services to heterogeneous users in the emerging iIoT, multi-tier computing was proposed by effectively combining edge computing and fog computing. More specifically, multi-tier computing systems compensate for cloud computing through task offloading and dispersing computing tasks to multi-tier nodes along the continuum from the cloud to things. In this paper, we investigate key techniques and directions for wireless communications and resource allocation approaches to enable task offloading in multi-tier computing systems. A multi-tier computing model, with its main functionality and optimization methods, is presented in details. We hope that this paper will serve as a valuable reference and guide to the theoretical, algorithmic, and systematic opportunities of multi-tier computing towards next-generation wireless networks.**

***Index Terms*—intelligent IoT, task offloading, multi-tier computing, resource allocation.**

I. Introduction

As the fifth generation wireless networks (5G) being commercially deployed, research efforts of the sixth generation wireless networks (6G) have begun to define 6G requirements and use cases. Four promising use cases have emerged. First, holographic telepresence allows realistic, full motion, three-dimensional (3D) images of people and objects to be projected as holograms into a meeting room to interact with each other in real time [1], [2]. Such remote holographic meeting, surgery, or distant learning will reduce the need for travel. The second key use case is digital twin, which creates a real-time, comprehensive, and detailed digital (virtual) copy of a physical object, or system [3]. Digital twins help push the boundaries of system reliability, used to support a wide range of capabilities such as diagnostics and fault prediction. The third one is connected industrial robots, such as Tactile Internet and intelligent cars. In this use case, the components of a control system (e.g., controllers, sensors, and actuators) are distributed across a wide geographic region [3], and therefore need to be connected via a wide area mobile infrastructure. In addition, these intelligent applications usually require stringent delay and jitter performance, with typical maximum tolerable network latency below 1 milliseconds. The fourth use case is automated network operation empowered by distributed artificial intelligence (AI), intelligent Internet of Things (iIoT), and big data technologies [4]–[6].

Many current and future applications require low latency, high reliability, and high data security protection [7]. These cannot be adequately met by the traditional cloud computing model, which requires to upload massive data and computing tasks to the cloud through fronthaul links and hence is difficult to meet the requirements of low latency and high energy efficiency. To provide low-latency services, a new computing paradigm called multi-tier computing was proposed by effectively combining edge computing and fog computing [8], [9]. With multi-tier computing, a large number of smart devices with varying computational resources, located around the end user, can communicate and cooperate with each other to execute computational tasks. A comparison between multi-tier computing and the current 5G-based edge computing is illustrated in Fig. 1. Multi-tier computing complements cloud computing and edge computing by offloading and dispersing computational (and communication and caching) tasks and resources along the continuum from the cloud to things.

Multi-tier computing makes the convergence of networking and computing possible by integrating with 5G and beyond systems [10], [11], supporting computational-intensive applications that require low latency but high energy efficiency, high reliability and high security, including a wide range of new novel applications such as augmented reality (AR), dynamic network slicing [12], Tactile Internet [13], industrial robots and intelligent robotic cars, smart grids, and smart cities, etc [14], [15]. The effectiveness of multi-tier computing depends largely on resource scheduling among edge and cloud nodes to reduce service latency and ease network congestion [9], [16]–[18]. Along the development of next-generation wireless networks, all kinds of user equipments (UEs) will be online all the time, promoting the advancement of iIoT and bringing diversified applications. These novel intelligent applications typically

K. Wang is with the Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China, and also with the School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China. (e-mail: klwang@cee.ecnu.edu.cn).

J. Jin is with the School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, Australia. (email: jiongjin@swin.edu.au).

Y. Yang is with Terminus Group, Beijing 100027, China, ShanghaiTech University, Shanghai 201210, China, and Peng Cheng Laboratory, Shenzhen 518055, China. (e-mail: dr.yangyang@terminusgroup.com).

T. Zhang is with the US National Institute of Standards and Technology (NIST). (e-mail: taozhang1@yahoo.com).

A. Nallanathan is with the School of Electronic Engineering and Computer Science at Queen Mary University of London, UK. (email: a.nallanathan@qmul.ac.uk).

C. Tellambura is with the Department of Electrical and Computer Engineering, University of Alberta, Canada. (e-mail: ct4@ualberta.ca).

B. Jabbari is with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, USA. (e-mail: bjabbari@gmu.edu).

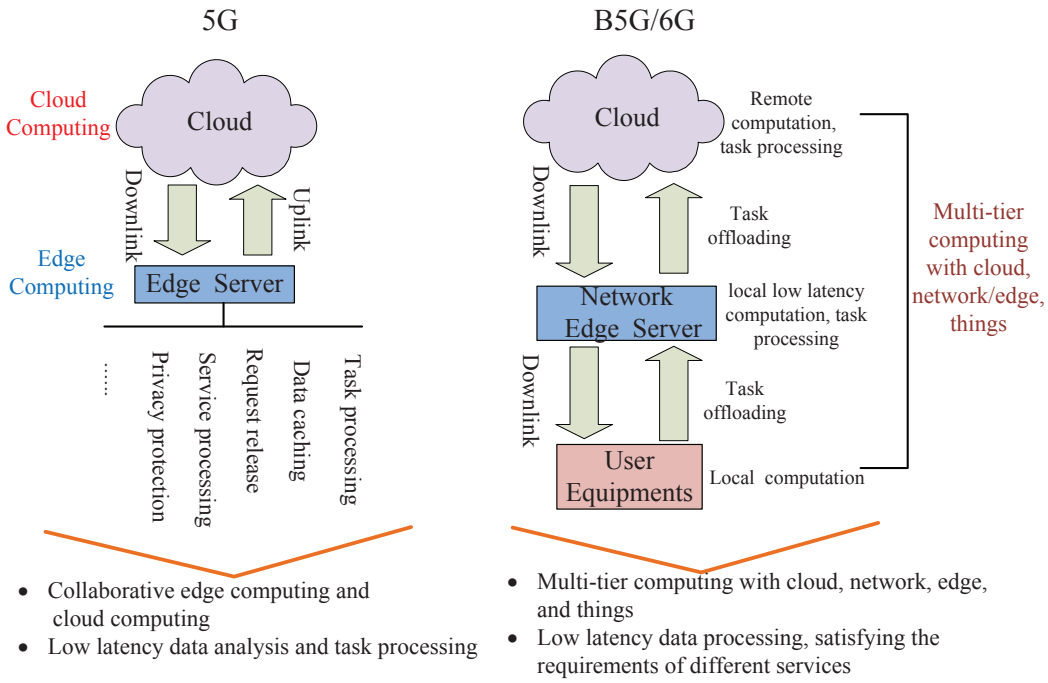

Fig. 1. Edge computing versus multi-tier computing

require low latency and demand prompt computations for real-time task processing and high data rates. However, mobile devices often have limited computation, storage, and energy resources. To overcome these limitations, it is essential to offload computational tasks from the end users to nodes in the multi-tier computing systems. Such task offloading enables distributed smart devices to share their idle computation and storage resources, facilitating the efficient utilization of multi-dimensional resources for low latency task processing. In multi-tier computing, fog/edge conducts task computation for delay sensitive applications at the network edge and cloud supports time-tolerant tasks via local task offloading systems. As a result, it realizes both real-time processing and local computational system control, which is crucial for not only robust control systems, but also for the low latency applications. Moreover, multi-tier computing systems will empower new task offloading models with the advancement of B5G and future 6G wireless communication system, as well as the new generation of embedded AI. As computational power moves from the cloud to edge and UEs, the computing and networking will be deeply integrated along the development of wireless communication systems. Therefore, cloud-to-things computing capabilities should be better coordinated, leading to a new stage of intelligent multi-tier computing systems.

A. Task Offloading in Multi-Tier Computing-based Next-Generation Wireless Networks

Next-generation wireless communication systems present various novel technologies, including massive multiple-input multiple-output (MIMO), intelligent reflecting surface (IRS), non-orthogonal multiple access (NOMA), millimeter-wave (mmWave) communications, space-air-ground integrated networks (SAGIN) and edge AI, etc. A multi-tier computing model integrates these radio technologies and AI to reduce task execution latency, allows large-scale user access, and enables efficient task offloading to realize efficient collaborative computing and multi-dimensional communication, caching, computation resource coordination. An example of multi-tier computing-based next-generation networks is illustrated in Fig. 2. Basically, it consists of two types of nodes, i.e., task node (TN) and helper node (HN). In particular, multiple TNs are able to offload their tasks to multiple HNs. It remains a fundamental challenge to effectively map multiple tasks or TNs into multiple HNs to minimize the total cost, such as task offloading latency or energy consumption, in a distributed manner, known as the multi-task multi-helper (MTMH) problem [19], [20].

Massive MIMO can provide array gains, diversity gains, and multiplexing gains without increasing spectrum and power resources. It has been shown in [21] that massive MIMO schemes improve significantly the data rates at the cell edge and also increase exponentially the spectrum efficiency, resulting in an order of magnitude increasing of system capacity. The integration of multi-tier computing and massive MIMO has been proven to enhance task offloading performance in terms of ultra reliability and low latency [17], [22]–[24]. In particular, Bursalioglu *et al.* [22] proposed an architecture of fog massive MIMO, and the system performance is analyzed by densely deploying a large number of multi-antenna base stations (BSs), where the users are served by zero-forcing beamforming (ZFBF). Wang *et al.* [17] proposed an energy-efficient task offloading framework in a massive multiple-input multiple-output (MIMO)-aided fog computing system, where

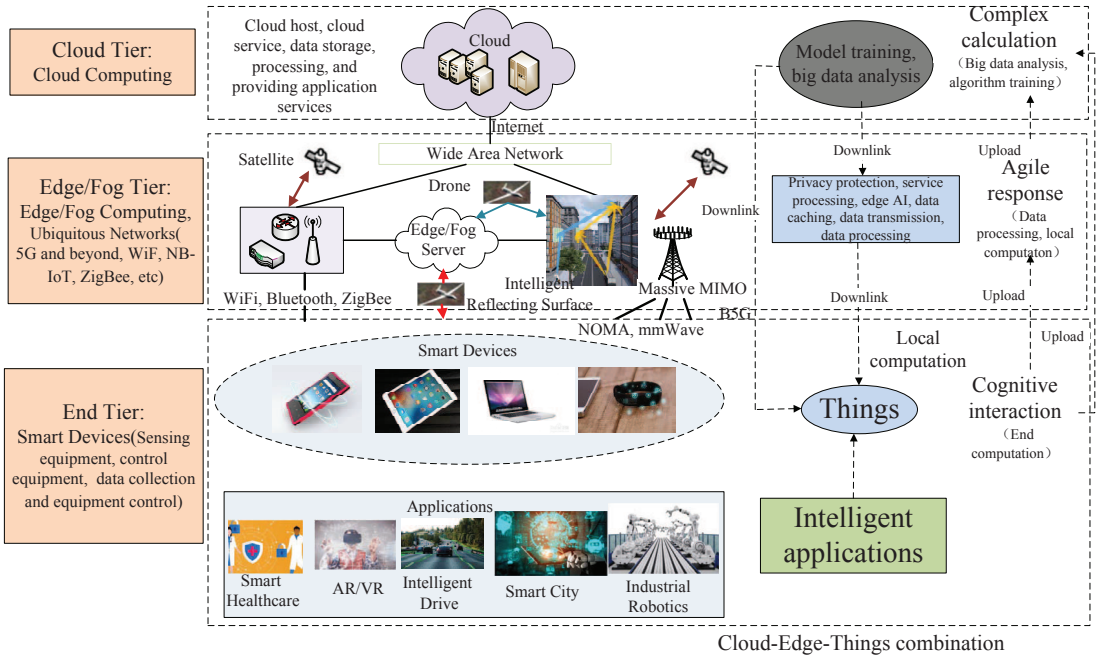

Fig. 2. Illustration of multi-tier computing network.

multiple task nodes offload their computational tasks via a massive MIMO-aided fog access node to multiple computing nodes for execution. In [23], Chen considered an edge computing framework based on distributed massive MIMO systems under fronthaul capacity constrain, aiming to minimize energy consumption on user devices. In [24], Mungara *et al.* proposed a new mechanism termed as dense fog massive MIMO, where the users are served by a large number of multiple antennas remote radio heads (RRHs), leading to high-throughput and low-latency transmission links. Although the above works demonstrate the advantages of massive MIMO-based multi-tier computing, the influence of imperfect channel condition on resource and task allocations is not studied, which is of paramount importance for time-varying multi-tier computing systems. On the other hand, to compensate for cloud computing, multi-tier computing systems provide computational capabilities both at the edge and center of the network. However, one of the major issues is how to manage task offloading and execution. More specifically, how to decide which tasks to perform at the end-user, fog/edge, or in the cloud. At a more granular level, the issue boils down to which node a particular task should be assigned to.

In B5G, the radio frequency may exceed 6 Gigahertz. Since higher-frequency signal is more sensitive to the blockage by obstacles, the coverage of each base station will be significantly reduced [25], [26]. Furthermore, devices at the cell edge or behind obstacles suffer from low task transmission rates, increasing both delay and energy consumption of task offloading in multi-tier computing systems [27]. IRS with a large number of low-cost reflecting elements, regarded as an effective auxiliary wireless communication technology for achieving high spectrum and energy efficiency, has attracted increasing attention to circumvent these restrictions and is listed as one of the candidate key technologies in 6G by academia and industry [28]–[34]. Thanks to the combination of array aperture gain (achieved by combining a direct transmission signal with an IRS reflection signal) and the reflection-assisted beamforming gain (achieved by controlling the phase shifts of IRS elements), IRS is able to improve the successful task offloading rate and the efficiency of resource scheduling in multi-tier computing systems. Therefore, IRS will be a key technology for task offloading in next-generation wireless networks. In [18], [35], the impact of IRS on computational performance is studied in a multi-tier computing system, demonstrating the benefits of the IRS to improve the task offloading, in comparison to the benchmark schemes.

Unlike orthogonal multiple access (OMA) techniques, non-orthogonal multiple access (NOMA) allows multiple nodes to concurrently communicate with a centre node over the same resource block, and hence enhancing the spectrum efficiency [36], [37]. Owing to the multiuser detection techniques such as successive interference cancellation (SIC) implemented at the receiver side [38], [39], NOMA can mitigate the co-channel interference, resulting in much better performance in terms of the network coverage and throughput compared to OMA techniques [40]. As expected, the integration of multi-tier computing and NOMA are able to boost the performance of multi-node task offloading [41]–[43]. Since higher-frequency signals like mmWave communications are highly correlated, making it conducive to the integration of NOMA, mmWave-NOMA is capable of supporting ultra-high bandwidth applications and massive access of users in multi-tier computing systems.

Meanwhile, by integrating satellite systems, aviation systems and ground communication systems, SAGIN is widely treated as a cornerstone of future 6G network. This new architecture supports seamless and near-instantaneous hyper-

connectivity [44], aiming at global data acquisition with high temporal and spatial resolution, high-precision real-time navigation and positioning, and broadband wireless communications. Being an essential component of SAGIN, UAVs are deployed flexibly at the air-network layer, assisting terrestrial network in task offloading and communications/computing/caching resources management due to their flexibility and proximity [45]. However, even with efficient task offloading, it is still not trivial to meet the quality of experience (QoE) requirements of heterogeneous users in the SAGIN.

Because of increasingly complex wireless networks, a typical 5G node is expected to have 2000 or more configurable parameters. Therefore, a recent new trend is to optimize task offloading and wireless resource allocation through AI technologies [46], [47], i.e., applying AI at multiple protocol layers (e.g., physical layer resource allocation, data link layer resource allocation, and traffic control) [48]. Thanks to the rapid development of mobile chipsets, the computational capabilities of edge devices have been substantially improved. For example, smart devices nowadays have as much computational capability as computing servers a decade ago. In addition, edge servers could provide end users with low latency AI services that are not possible to achieve directly on the devices. Since the computational resources of edge servers are not as much as those of cloud centers, it is necessary to adopt joint design principles across edge servers and edge devices to reduce task execution latency and enforce privacy for task offloading [46]. As a result, advances in multi-tier computing systems offer an opportunity to move the frontiers of AI from the cloud center to the edge of the network, inspiring a new field of research called edge AI, including both AI model training and inference procedures.

In order to realize low-latency task processing and provision computing, storage and networking services, distributed AI and federated learning algorithms are performed on multi-tier computing servers at the access network [49]. Wireless networks with AI can support the on-demand intelligent low-latency services, commencing to emerge in complicated wireless network management and resource optimization [50], [51]. Since the data are processed at the edge server in the close proximity of smart device, there is no need to transfer a large amount of raw data to the back-end server. Thus, using edge AI on task offloading infrastructure not only saves network bandwidth on backhaul links, but also reduces greatly the task execution latency. Edge AI will be a significant step towards reducing task execution latency by intelligently enabling task offloading and local caching of popular file and content migration. In addition, intelligent task offloading for computational tasks will make it possible to further virtualize users' handsets and improve battery lifetime. In all, edge AI provides a new paradigm of optimization algorithms design for efficient task offloading and service-driven resource allocation in multi-tier computing systems [11]. By seamlessly integrating sensing, communications, computing and intelligence, edge AI will empower multi-tier computing systems to support multiple intelligent applications, including industrial robots, intelligent robotic cars, and intelligent healthcare etc.

B. Main Contributions

Although the above discussions have demonstrated the benefits of task offloading in wireless communication systems, to the best of our knowledge, the task offloading with multi-tier computing resources in next-generation wireless networks has not been well studied. In this paper, a vision of multi-tier computing with intelligent task offloading is presented, focusing on its interactions with various wireless techniques and resource allocations. Future research directions and open problems are then discussed, embracing the era of multi-tier computing based next-generation wireless networks.

Against the above backdrop, our contributions could be further detailed as follows:

- The vision, challenges and solutions for task offloading in multi-tier computing systems towards next-generation wireless networks.
- The task offloading in multi-tier computing systems is presented, including the massive MIMO-aided task offloading, the task offloading with IRS, the task offloading with NOMA and mmWave, the task offloading in Space-Air-Ground Integrated Networks (SAGIN), and edge AI-empowered task offloading.
- The multi-tier computing resource allocation for task offloading is elaborated. Specifically, we introduce the main functionality and optimization methods as well as the algorithms for task offloading in multi-tier computing systems.
- We discuss the research directions and open problems of task offloading for multi-tier computing-based next-generation wireless networks.

C. Paper Organization

The rest of the paper is organized as follows. Section II introduces the enablement of multi-tier computing for next-generation wireless networks, while Section III presents the resource allocation for multi-tier computing systems. Section IV is focused on research directions and open problems for multi-tier computing. In Section V, we provide our conclusions.

II. Enablement of Task Offloading for Multi-Tier Computing-based Next-Generation Networks

In this section, we present the vision, challenges and solutions for task offloading in multi-tier computing systems, including the massive MIMO-aided task offloading, the task offloading with IRS, the task offloading with NOMA and mmWave, the task offloading in SAGIN, and edge AI-empowered task offloading.

A. Massive MIMO-Aided Task Offloading

With the advent of next-generation of wireless standards, new high-performance technologies are introduced. One of these key technologies is massive MIMO [52] that has been increasingly adopted in different networking and computing frameworks. However, the works of [9], [53]–[55] mainly considered single-antenna computation offloading systems,

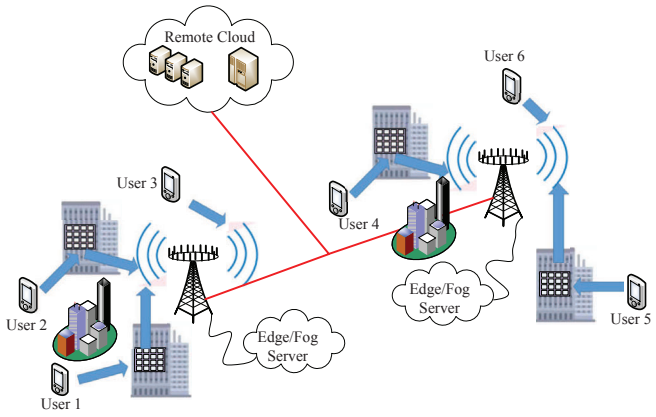

Fig. 3. Illustration of a massive MIMO and IRS-enabled multi-tier computing network.

by taking joint resources allocation and task offloading into account, but failed to exploit the MIMO advantages for task offloading efficiency. As we know that MIMO techniques have the potential of achieving high spectral efficiency (SE) [56]–[58], they have been introduced to boost the performance of edge users by increasing the task offloading data rate. In particular, equipping the base stations (BSs) with a large number of antennas, widely known as massive MIMO, has emerged as one of the most promising solutions [59], [60] to significantly improve system SE and energy efficiency trade-off. More specifically, as the number of antennas increases, channels become more deterministic, known as channel hardening. Data rates and communication resource allocations are hence largely determined by large-scale fading. This implies that resource allocation does not need to be updated frequently, leading to significant savings in signal transmission overhead. In summary, massive MIMO schemes improve spectrum and energy efficiency and support an increased number of users, both of which are critical for multi-tier computing systems.

As the core technology of wireless communication, relay technique has been integrated into various wireless communication standards to improve network coverage and throughput [61]. In particular, massive MIMO-enabled relay networks can enhance spectral efficiency and achieve more reliable data transmission for spatially distributed user nodes through intermediate massive antenna relay nodes [62], [63]. Thus, a massive MIMO-aided fog access node (FAN) serving as a relay is capable of significantly improving the data rate of offloaded tasks and the task execution efficiency. The new computing model that combines massive MIMO with multi-tier computing will facilitate efficient task offloading of computation-intensive tasks to achieve efficient collaborative computing and multi-dimensional communication, caching, computation resource scheduling.

B. Task Offloading with IRS

Next, we will introduce a concrete example of implementing IRS in multi-tier computing systems to reduce task offloading latency and energy consumption, as shown in Fig. 3. Each user could either offload its task to the multi-tier nodes such as edge/fog server for computation via the IRS or to the cloud via the IRS and massive MIMO node. In order to further improve uplink task offloading performance for resource-limited end users, IRS technology has attracted extensive attention due to its advantages of low cost, easy deployment, fine-grained passive beamforming, and directional signal enhancement or interference nulling. By controlling surface reflective elements, IRS can be reconfigured to provide a more favorable wireless propagation environment for communications. Obviously, using IRS in multi-tier computing systems is an economical and environmentally friendly method to facilitate task offloading [18].

In [35], Chu *et al.* studied the impact of an IRS on computational performance in a mobile edge computing (MEC) system, targeting to optimize the sum computational bits and taking into account the CPU frequency, the offloading time allocation, transmit power of each device as well as the phase shifts of the IRS. In [18], Wang *et al.* investigated the task offloading problem in a hybrid IRS and massive MIMO relay assisted fog computing system, and formulated a joint task offloading, IRS phase shift optimization, and power allocation problem to minimize the total energy consumption. In [64], Zhou *et al.* studied an IRS-assisted MEC systems, in which IRS is deployed to assist task offloading from two users to the fog/edge access point connected to the edge cloud. Under the constraint of IRS discrete phase, the passive reflection phase of IRS and the user's computational task scheduling strategy is designed to minimize the total task processing latency. In [65], Bai *et al.* studied an innovative framework to employ IRS in wireless powered MEC systems, and the task offloading is based on orthogonal frequency-division multiplexing (OFDM) systems. The objective is to minimize the total task offloading energy consumption. On the basis of the above studies, it is evident that IRS can provide an additional link both for data transmission and for task offloading, so as to enhance computational capability.

C. Task Offloading with NOMA and mmWave

As we all know, NOMA performs significantly better in terms of the network coverage and spectrum efficiency than OMA [40]. Under this circumstance, the integration of multi-tier computing and NOMA is able to achieve far better performance for task offloading compared to multi-tier computing with OMA [41]–[43]. In particular, Wang *et al.* in [43] proposed a NOMA-based fog computing framework for industrial Internet of Things systems, where multiple task nodes offload their tasks via the NOMA strategy to multiple computing nodes for task computation. Accordingly, they formulated a joint task offloading and subcarrier allocation problem to minimize the total cost in terms of energy consumption and latency subject to the given communication and task computation constraints. In addition, Zhang *et al.* [41] proposed a network architecture of NOMA-based Fog Radio Access Networks (F-RANs), where the resource allocation with power and sub-channel allocation is studied to improve the network performance. Moreover, Wen *et al.* [42] and Wang *et al.* [66]

formulated an energy efficiency maximization and a task completion time minimization problem in NOMA-enabled fog/edge computing networks, respectively. All the above works have manifested that the NOMA-based task offloading scheme can significantly reduce the energy consumption and latency cost compared to its OMA counterpart.

Regrading task offloading with mmWave, Zhao *et al.* have proposed single-user and multi-user mmWave task offloading framework respectively in [67] and [68], both of which aim to minimize the task offloading latency exploiting the benefits of the mmWave communications. As mentioned before, mmWave is conducive to the integration of NOMA. Yu *et al.* [69] analyzed the impact of NOMA and mmWave on task offloading, where the hybrid beamforming at the BS and the resource allocation at the end user are jointly optimized, and NOMA-mmWave is shown to improve the computation efficiency by the theoretical and simulation results.

D. Task Offloading in SAGIN

IoT seeks to connect billions of resource-constrained devices around us through heterogeneous networks. The SAGIN is viewed as a major candidate to support such IoT requirements, provisioning seamless and massive connectivity for smart services [44], [70]. In the past two years, 5G wireless networks have been commercialized and deployed around the world. Although 5G is still in its development, academia and industry have now shifted their attention to beyond 5G and 6G wireless networks, in order to meet the demands of ultra-low latency and high energy efficiency for iIoT [44]. Among the discussions about 6G, from the perspective of computing, communication, and caching, it is the trend to combine SAGIN with multi-tier computing technologies in the 6G networks.

Specifically, it is widely recognized that SAGIN will be the potential core architecture of the future 6G network to support seamless and near-instantaneous hyper-connectivity [44]. Thus, multi-tier computing with SAGIN promotes the task offloading performance. As a key part of this, in the integrated air-ground branch, unmanned aerial vehicles (UAVs) are flexibly deployed at the aerial network layer, assisting in communication, computing and caching of ground networks due to their flexibility and proximity [45]. However, in 6G networks, SAGIN still faces challenges such as the demands of temporal-spatial dynamic communication/computing/caching services, large-scale complex connection decisions and resource scheduling, and ubiquitous intelligence demands within the network. To sum up, it remains extremely challenging to realize these visions of 6G in SAGIN.

There have been heavy research efforts on the architecture of SAGIN and multi-tier computing in the existing literature. Cheng *et al.* [45] proposed a novel air-ground integrated mobile edge network, by investigating the potential benefits and applications of drone cells, and UAV-assisted edge computing and caching. To support diverse vehicular services, Zhang *et al.* [71] presented a software defined networking (SDN)-based space-air-ground integrated network architecture. Focusing on provisioning computing services by UAVs, Zhou *et al.* [72] proposed an air-ground integrated MEC framework to cater for the urgent computing service demand from the IoTs. Furthermore, Kato *et al.* [73] conducted a comprehensive study about how to deal with the challenges related to the space-air-ground integrated networks by AI techniques, including network control, spectrum management, energy management, routing and handover management, and security guarantee. In [74], Cheng *et al.* demonstrated a SAGIN edge/cloud computing architecture for offloading the computation-intensive applications, considering remote energy and computation constraints, and developed a joint resource allocation and task scheduling approach to efficiently allocate the computing resources. In [75], Shang *et al.* studied MEC in air-ground integrated wireless networks to minimize the total energy consumption by jointly optimizing users association for computation offloading, uplink transmit power, allocated bandwidth, computation capacity, and UAV 3-D placement. However, how the air network layer allocate the communication/computing/caching resources intelligently for task offloading of the ground network layer in SAGIN has not been adequately addressed.

E. Edge Intelligence-Empowered Task Offloading

With the continuing increase in the quantity and quality of rich multimedia services, the traffic and computational tasks of mobile users and smart devices have significantly increased in recent years, bringing huge workload to the already congested backbone and access networks. Even with the help of multi-tier computing systems, it is challenging to satisfy the quality of experience (QoE) requirements of users. The main difficulty lies in the need of large amount of wireless data and task transmissions for task offloading, causing wireless channel congestion. Therefore, the optimization problem or decision making of the combined wireless communication resource allocation and multi-tier task offloading is the key. That is, how to share the communication resources and computing resources between edge nodes and the cloud. In response to the increasing complexity of wireless communication networks, AI technologies have been proposed as a new research trend to optimize resource allocations [46], [47], including but not limited to applying AI algorithms to physical layer resource allocation, data link layer resource allocation, medium access control, and traffic and congestion control [48]. Especially, reinforcement learning is often applied to jointly manage communication, computing, and caching resources. With learning based multi-tier computing systems, we can optimize task offloading, communication resource allocation, and content caching at edge nodes. Further, federated learning [76], as a distributed learning framework, always brings the following benefits for task offloading: 1) great reduction of the amount of data that must be uploaded through wireless uplink channel, 2) cognitive response to the changing wireless network environments and conditions, and 3) strong adaptability to the heterogeneous nodes in the wireless networks, 4) better protection of personal data privacy.

In learning-based multi-tier computing systems, task offloading decision and communication resource allocation vectors generally are binary variables, turning out challenging to

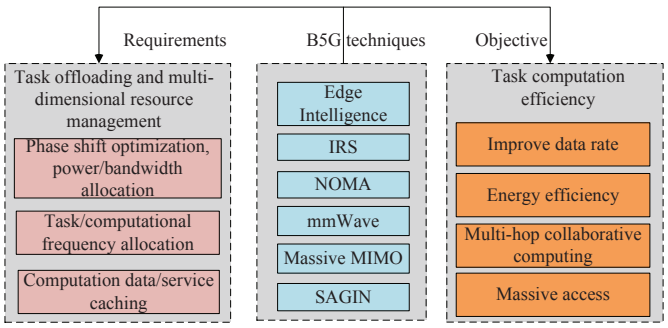

Fig. 4. Requirements and objective for designing multi-tier computing systems.

find the optimal solution of resource allocations. Moreover, the feasible set and the objective function of the optimization problem are generally nonconvex, making the problem NP hard. In addition, in time-variant systems, channel conditions and computational cost are dynamic. Instead of solving the NP hard optimization problem by utilizing conventional optimization methods, the task offloading and communication resource allocation problem in multi-tier computing systems could be possibly solved using online learning algorithms. During the online learning process, the deep reinforcement learning methods might be applied to jointly optimize the subcarrier allocation and task offloading in each time episode. Online federated learning framework is recently utilized to learn in a distributed way, in order to solve the task offloading and communication resource allocation problem. Based on the communication, computational resource allocation, the multi-tier task offloading decisions can then be optimized.

Furthermore, edge learning methods have been investigated in some edge/fog computing systems to simplify the optimization algorithm or fulfill online implementations [16], [43], [77]–[82]. In [78], Huang *et al.* designed a deep learning-based task offloading strategy to minimize weighted energy consumption and latency. In [81], Wang *et al.* leveraged deep reinforcement learning method for smart resource allocation in a software defined network (SDN)-enabled edge computing architecture. In [80], Huang *et al.* proposed a deep learning-based task offloading strategy for offloading decisions and resource allocation of a wireless powered edge computing system. In [83], Yang *et al.* also used deep reinforcement learning method in IRS-aided edge computing systems to enhance system security and maximize the sum rate of the down-link task offloading. In [84], a convolutional neural network was constructed for channel estimation of a large IRS-aided massive MIMO communication system to estimate the direct and the cascaded channels, used for multi-tier task offloading.

III. Multi-tier Computing Resources Allocation for Task Offloading

In this section, we first characterize the multi-tier computing resource allocation in next generation wireless networks, and then effective optimization methods are presented to achieve efficient task offloading with multi-tier resources.

A. Main Functionality

In this subsection, the computational and communication resources allocation, service placement, and security requirement are characterized for designing multi-tier computing systems, which is illustrated in Fig. 4.

1) Computation: Multi-tier computing architectures were envisioned to achieve rapid and affordable scalability by developing computation capabilities flexibly along the entire cloud-to-things continuum [85]. In essence, multi-tier computing systems distribute computing capability anywhere between the cloud and the things to take full advantage of the computational resource available along this continuum, thus extending the traditional cloud computing architecture to the edge of the network. Thanks to multi-tier computing, some application components can be performed at the network's edge, like delay-sensitive components. While other components, such as time-tolerant and computation-intensive ones, are performed in the cloud. Satisfying diverse delay requirements will require both cloud computing with enormous resources to support time-tolerant tasks, and distributed fog/edge computing with limited resources and simple algorithms closer to the users to support time-sensitive tasks. With heterogeneous computing resources and collaborative service architecture, the proposed multi-tier computing systems are able to effectively support a full range of services in different environments. On this basis, multi-tier computing provides the advantage of low-latency task offloading since it allows task to be processed at the network edge, close to the end devices. Obviously, cloud computing alone is not adequate for supporting all IoT applications, while a multi-tier computing system can be complementary.

For smart devices with abundant computing resources, multi-tier computing seeks to achieve seamless integration of edge and cloud systems. This vision goes beyond treating the network edge and smart devices as separate computing platforms. Seamlessly integrating fleets and swarms of mobile IoT entities into a dense multi-tier enclave is a new distributed computing paradigm that improves the scalability, extensibility and assemblability of cloud services through edge of computing systems. Smart devices (cars, drones and robots) have spare computational resource, allowing the multi-tier computing platform to reduce energy consumption and task processing latency compared to the traditional edge computing scenarios relying on static and low-power edge servers.

2) Communication: Multi-tier computing systems distribute communication functions anywhere between the cloud and things to take full advantage of the communication resource available along this continuum. In massive MIMO-aided multi-tier computing systems, the achievable data rates are mostly determined by large-scale fading, and so is the communication resource allocation. This means that there is no need to frequently update communication resource allocations, hence reducing signaling overhead. IRS is capable of improving the success rate of the task offloading. Given the

potential gains, if the line-of-sight (LoS) link between the task offloading nodes and computing nodes is blocked by obstacles, the task could be offloaded via the IRS reflected links. In this manner, we attempt to optimize the link selection and wireless communication resource allocation.

It is important to maintain the required data rates for task offloading. Take the task offloading from a car as an example. Given that a connected car produces tens of megabytes of data per second, an autonomous vehicle may generate up to a gigabyte per second [8]. Here, dense moving edge nodes can support accelerated data communication by largely utilizing directional high-rate communication in the massive MIMO, IRS or the SAGIN. Edge nodes at the same time provide novel strategies for smart devices to combine the benefits of centralized and ad-hoc topologies into a unified solution by using multi-hop, multi-connection mechanisms to communicate with adjacent network infrastructure when facing the intermittent connectivity.

3) Storage: Because the edge nodes often have limited storage resources, distributing the data among edge and cloud nodes is vital for optimizing task offloading latency or energy consumption at a given QoS level. On top of it, multi-tier computing also brings a large amount of cloud-like services closer to the end users. Caching computational data or services at edge nodes is hence crucial, which relieves the burden of backhaul transmission with transporting all the data to the clouds.

Accordingly, elastic storage capacity of edge nodes might be used to support applications running on resource constrained IoT devices. Due to the inherent flexibility of multi-tier computing systems, it is possible to integrate a large number of densely distributed devices. Caching capacity of edge servers is usually accessed by both smart devices and edge access points. For example, user nodes are possibly consolidated into special capacity areas. Then, multiple interconnected edge infrastructures that coexist in space and time could pool storage resources of adjacent edge networks together for sharing by smart devices and end users.

4) Security: In cloud computing systems, massive data need to be uploaded to the cloud data center through a fronthaul link, where data security cannot be guaranteed. However, multi-tier computing systems present unique security challenges and opportunities. Dense edge nodes with established dynamic trust chains are acting as a trusted authority for other smart devices and systems. In particular, multi-tier computing systems with edge and cloud can handle responsibilities such as trusted computing platforms, and secure storage of short-term sensitive information. Multi-tier computing systems also utilize edge systems to facilitate local threat monitoring, detection, and protection for users and provide powerful proximity-based authentication services for better authentication through proxy smart devices.

However, the multi-tier computing systems meanwhile incur new security vulnerabilities, mainly from multiple heterogeneous nodes. For example, in a multi-node environment, when multiple potentially competing service providers and consumers share resources distributed across a set of hardware platforms, advanced authorization and authentication mechanisms should be created to effectively leverage this heterogeneous medium and devices between edge and cloud entities. Fortunately, a trusted execution environment supported by a public key infrastructure may be a suitable solution to the above problems. Nevertheless, the intelligent integration of hardware assistance and software security mechanisms in multi-tier computing systems remains an open research question.

Additionally, multi-tier computing systems have to cope with changing environments compared to existing edge computing systems that mainly operate under known conditions. In this case, the security mechanism for multi-tier computing systems is supposed to constantly adapt to the changing operating conditions. To address this challenge, multi-tier computing systems must dynamically adjust their overall security posture. It requires the design of new security protocols, which is able to respond to any security threat without causing service disruptions and to fulfill secured and uninterrupted operation of the task offloading.

B. Optimization Algorithms

With the exponential growth of real-time services, delay has emerged as a key figure of merits and become the design metric of multi-tier computing systems [86], and the total task offloading latency consists of the task computation latency at multi-tier computing nodes plus the round-trip task transmission latency. For local computing, the latency only includes the processing latency of local CPUs. For task offloading, if a task is to be processed by the edge/fog or cloud, the node needs to transmit the task through the shared wireless channel. Hence, the latency includes the task transmission latency and task computation latency in the edge/fog or cloud. In the meantime, IoT services urge for more and more computation and communication resources due to the rapid increasing number of connected devices. However, as these intelligent devices usually have limited computation and energy resources, it is a big challenge for the service providers to promote these novel applications. In this subsection, the effective optimization methods are presented to achieve low latency and energy efficient task offloading with multi-tier resources, and to decide which tasks to perform at the end-user, fog/edge, or in the cloud with dynamic resources. Therefore, the differences and difficulties in multi-tier computing systems need to solve a series of non-convex optimization problems with binary variables, as well as the stochastic variables.

1) Nonconvex Optimization: During task offloading process, most of the resource allocation problems in multi-tier computing systems need to solve a series of nonconvex optimization problems. For example, for IRS-aided multi-tier computing systems, there are four blocks of optimization variables, namely, task offloading ratio, power allocation at the relay node, and IRS phase shifts of two hops' task transmission. The optimization of task offloading ratio is related to the computing setting, while the optimization of power allocation and phase-shift matrices affects the communication design. However, the resource allocation problem in IRS-enabled multi-tier computing systems is difficult to solve due

to two aspects. The first one is the coupling effect between the power allocation vector and the IRS phase-shift vector. The second one is that the objective function is non-convex with respect to the phase shifts. Obviously, it is an open challenge to obtain a globally optimal solution directly. In fact, alternating optimization technique is a widely applicable and efficient approach for solving optimization problems involving coupled optimization variables, which has been successfully applied to several communication resource allocation problems such as hybrid precoding [87], power allocation [88], and IRS phase shift optimization [28], [89]. In this case, a locally optimal solution is usually provided. To be specific, the resource allocation optimization problem can be transformed into a phase shift optimization problem, a power allocation problem, and a task allocation problem, respectively, by using the popular alternate optimization technique to decouple communication and computational design.

Remarkably, in contrast to the alternate optimization technique, distributed optimization algorithms for non-convex optimization have appeared in the literature [90]. In [91], [92], Tatarenko *et al.* and Zeng *et al.* studied distributed gradient descent methods for unconstrained non-convex optimization problems, respectively. Distributed optimization algorithms are generally divided into two categories: discrete time algorithms and continuous time algorithms. The existing work mainly focuses on discrete time algorithms, while continuous time problem has attracted extensive attention in recent years, mainly because of the wide application of continuous time setting in practical systems and the development of continuous time control technology. In addition, discrete time and continuous time algorithms are closely related to each other due to the time scale transformation. Specifically, when the time step size approaches zero, the optimization algorithm for discrete time system is similar to the continuous one. Note that coupled non-linear constraints are also an important constraint in distributed optimization problems. However, distributed algorithms dealing with coupled non-linear constraints are basically convex problems, i.e., both the objective function and the constraint are convex. In [86], Wang *et al.* developed distributed augmented Lagrangian based algorithms for non-convex optimization problems of multi-tier computing networks subject to local constraints and coupled non-linear equality constraints, and investigated the joint design of the task offloading, service caching and power allocation to minimize the total task scheduling delay.

2) Mixed-Combinatorial Optimization: Combinatorial optimization problems have been analyzed in many works (e.g., [93]–[96]). Under the framework of combinatorial optimization, an important trend is analyzing combinatorial optimization problem within the framework of Euclidean combinatorial optimization, whose optimization is carried out in a Euclidean space. In [95], [96], Barbolina *et al.* and Yemets *et al.* studied the Euclidean combinatorial optimization problems, and investigated the properties of its convex hull and methods of solving separate classes of Euclidean problems of combinatorial optimization. Additionally, the general permutation set problem is an important Euclidean combinatorial optimization problem.

As previously mentioned, the resource allocation problems in multi-tier computing systems involve optimizing computation, communication and caching. In general, task offloading, task data caching and communication resource allocations are binary variables. Specifically, in multi-tier computing for next-generation wireless networks, we need to jointly optimize the subcarrier and bandwidth allocation [97]–[99], transmit power and receive beamforming [17], [43], passive beamforming at IRS [18], device selection [9], [18], location updates task offloading [16], and computational frequency control [9], so as to reduce the latency and energy consumption in the task offloading procedure. Therefore, these resource allocation schemes can be formulated as a mixed combinatorial optimization problem that requires joint optimization of continuous value variables (e.g., beamforming, power control) and discrete value variables (e.g., task allocation, service placement, subcarrier allocation).

It should be noted that the existing optimization methods for mixed combinatorial optimization problems are mainly based on traditional iterative optimization approaches [18], [100]–[103], or adopt a direct end-to-end online learning approaches [16], [81]. However, they may not achieve good trade-off between algorithm complexity and resource allocation performance. Additionally, reinforcement learning (RL)-based approaches are often involved to solve combinatorial optimization problems that are unconstrained or have few constraints due to feasibility issues [43], [104]. Deep RL requires a Markov process to achieve satisfactory resource allocation performance [105]. However, Markov process may not exist in practical combinatorial optimization problems, as they have many non-convex constraints with memory. This results in difficult design of reward features for Markov optimization process, unfeasible solutions, and potential degradation of overall performance.

3) Stochastic Optimization: In multi-tier computing systems, stochastic optimization approach only relies on the probabilistic description about the uncertainty of computation capacity and radio channel condition, and is able to provide a trade-off between conservatism and probabilistic assurance for the achievable task offloading performance. Stochastic programming has been widely studied in the past decade due to its wide application in machine learning and resource allocation. In a stochastic optimization problem, the objective function or constraints are the expectation of some function of random variables (such as estimated computation capacity and channel condition in learning approach) [106]. The challenge of stochastic optimization is that the distribution of the random variables is often unknown. Most existing literature on stochastic programming assumes that the basic distribution of random variables is fixed and that independent samples are sequentially drawn from this common distribution. However, the basic distribution of random variables involved in stochastic optimization may change slowly over time in many practical applications.

Stochastic optimization in state-based systems with discrete or continuous time are often modeled as Markov chains. Their effective optimization method is an important research topic. The Markov model has a wide range of applications, especially

in the area of task offloading in multi-tier computing systems. Specifically, some work modeled the task offloading problem as a stochastic programming problem, and jointly optimized the task allocation and the communication resources allocation [107]. However, in all these works, system parameters need to be acquired offline, which is impossible for a time-varying system [108]. It should be noted that there are multiple dynamic parameters in multi-tier computing systems. Therein, user mobility and channel condition are intrinsic features of wireless networks when nodes are usually in a mobility state. Then, due to changes in network topology, these parameters are time-varying, and the stochastic task offloading framework is considered as a method of online learning where users can learn time-varying system parameters.

In many multi-tier computing applications, optimization criteria are trade-offs between several competing goals, such as computational cost minimization and profit maximization. In this tradeoff model, it is important to establish an optimal strategy that may often not be intuitive. However, there are also optimization problems with no tradeoff characteristics, leading to counterintuitive optimal strategies. Therefore, the use of Markov decision process (MDP) to optimize stochastic systems should not be ignored. A reinforcement learning task that satisfies the Markov property is called an MDP, which is a tool for modeling sequential decision-making problems. If the state and action spaces are finite, then it is called an finite MDP, which is crucial to the basis of reinforcement learning (RL). In MDP, future decisions are based on recent state. Thus, an optimal policy consists of actions based on the observed history to maximize the expected reward. RL has been widely adopted in the unknown environment, by continuously interacting with the environment to achieve the optimal results out of imperfect information. RL can further take advantage of high-dimensional characteristics of deep neural network (DNN), evolving into deep reinforcement learning (DRL). It is worthwhile noticing that DRL is capable of characterizing the infinite states caused by measurement error and environmental noise. There exists extensive literature in online learning task offloading for stochastic optimization, such as DRL [109]–[112], which generally target at a broader set of learning problems in MDPs. As far as we are aware, high-dimensional action spaces are still an urgent and challenging problem in DRL. To make the problem tractable, the general optimization problem is reduced into an MDP that only considers a meaningful parameter. Furthermore, the Multi-Armed Bandit (MAB) problem is a special case of MDP problems for which regret learning frameworks are generally considered to be more efficient of computational complexity. Additionally, the use of the MAB model is appropriate and recognizable, taking advantage of the fact that the resources of edge node are limited. Based on the above analysis, MDP promises an online learning framework for learning computing resources and available communication, storage resources information for stochastic optimization, aiming to minimize task offloading cost.

IV. Research Directions and Open Problems

In this section, we present the research directions and open problems for task offloading in next-generation wireless networks, supported by the wireless network infrastructures in Section II.

A. Multi-Dimensional Resource Management

Compared to cloud computing, the edge nodes and end users in multi-tier computing systems may have limited resources. Therefore, communication, computing and caching resource allocation is a very important research issue in multi-tier computing systems. Specifically, next-generation wireless communication networks present various technologies, including massive MIMO, IRS, NOMA, SAGIN, and AI etc. These new communication technologies integrated with multi-tier computing will reduce task offloading delay, grant large-scale user access and promote rapid development of the intelligent services, as well as realize efficient collaborative computing and multi-dimensional communication, caching, computation resource sharing through efficient task offloading.

However, the computation power of multi-tier servers is typically limited. The wireless physical layer resource allocation and user access techniques are the key challenges that hinder the success of multi-tier computing for 5G and beyond in executing compute-intensive and latency-critical applications. The optimization of resource allocation may be multi-objective in different situations, e.g., diverse nature of applications, heterogeneous server capabilities, user demands and characteristics, and channel connection qualities.

B. Multi-Tier Task Allocation

Since multi-tier computing systems provide extra computing capability at the network edge, one of the core problems is how to manage task allocation. More specifically, how to decide which tasks should be performed on end-user devices, at fog/edge systems, or in the cloud. At a more granular level, the challenge is to which computing nodes should a task be assigned. To achieve low latency and high energy efficiency of task offloading, computing tasks need to be scheduled to computing nodes with different capabilities according to different task computing models, communication bandwidths and channel qualities. Therefore, heterogeneity becomes an important factor in multi-tier computing architectural design. Dealing with different task computation and various communication protocols to manage task offloading becomes a major problem.

C. Heterogeneous QoS Management

With the development of various novel technologies, intelligent services are increasingly applied in many fields of human life, including business, manufacturing, health-care, entertainment, etc. On one hand, the number of smart services deployed around edge and cloud servers is growing rapidly. On the other hand, different service providers provision services with similar functions, and different edge servers may possess different service performance. Then, the smart devices will

require services with different QoS requirements. In light of these descriptions, intelligent services are migrating to the network, i.e., to edge servers residing near end users.

Note that QoS requirements in multi-tier computing systems include task response time, throughput, reliability and availability, typically different for different users. However, user mobility and different server capabilities turn the applicability of traditional QoS management inapplicable. Therefore, how to monitor and manage QoS attributes, and schedule multi-dimensional resources timely and effectively to fulfill specific QoS requirement for each user becomes the main issue in multi-tier computing systems.

D. Data Privacy

In multi-tier computing systems, data and computation task need to be collected close to the physically distributed edge devices, and there exists a large number of devices in the systems. When analyzing sensitive information from distributed nodes, data privacy cannot be compromised. We should select computing nodes in a way that best protect the data privacy, considering the computing nodes in different parts of the network may have different privacy protection capabilities. The tasks collected, transmitted, and processed at the edge or in the cloud should be anonymized [113]. Then, multi-tier data analysis and processing is achieved securely in multi-tier computing systems. Note that distributed systems are in general more vulnerable to be attacked than centralized systems, and both the end devices and edge computing nodes in multi-tier computing systems are typically less powerful than the cloud. Therefore, these nodes may not have as adequate resources as the cloud to protect themselves. In addition, the devices and edge computing nodes may not have enough intelligence and capability equipped to detect threats due to limited resources. In all, data privacy of multi-tier computing from things to the cloud will be the focus of future research in multi-tier computing systems.

V. Conclusions

In this paper, we investigated the key wireless communication techniques, effective resource allocation approaches and research directions to embrace the era of task offloading for multi-tier computing-based next-generation wireless networks. In particular, the multi-tier computing system model, multi-tier computing resources and optimization methods were presented for better facilitating the task offloading. We hope that this paper will serve as a valuable reference and guide to further promote the theoretical, algorithmic, and systematic development and advancement of task offloading with multi-tier computing resources in next-generation wireless networks.

References

[1] K. David and H. Berndt, "6G vision and requirements," *IEEE Veh. Technol. Mag.*, vol. 13, no.9, pp. 72–80, Sep. 2018.

[2] A. H. Khan, N. U. Hassan, C. Yuen, J. Zhao, D. Niyato, Y. Zhang, and H. V. Poor, "Blockchain and 6G: The future of secure and ubiquitous communication," *IEEE Wireless Commun.*, pp. 1–8, Jan. 2022.

[3] B. Zong, C. Fan, X. Wang, X. Duan, B. Wang, and J. Wang, "6G technologies: Key drivers, core requirements, system architectures, and enabling technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no.3, pp. 18–27, Sep. 2019.

[4] K. Yang, Y. Shi, Y. Zhou, Z. Yang, L. Fu, and W. Chen, "Federated machine learning for intelligent IoT via reconfigurable intelligent surface," *IEEE Network*, vol. 34, no.5, pp. 16–22, Sep. 2020.

[5] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang, and X. S. Shen, "Software defined space-air-ground integrated vehicular networks: Challenges and solutions," *IEEE Commun. Mag.*, vol. 55, no.7, pp. 101–109, July 2017.

[6] L. Lyu, J. C. Bezdek, J. Jin, and Y. Yang, "FORESEEN: Towards differentially private deep inference for intelligent internet of things," *IEEE J. Select. Areas Commun*, vol. 38, no.10, pp. 2418–2429, Oct. 2020.

[7] S. Andreev, V. Petrov, K. Huang, M. A. Lema, and M. Dohler, "Dense moving fog for intelligent IoT: Key challenges and opportunities," *IEEE Commun. Mag.*, vol. 57, no.5, pp. 34–41, May 2019.

[8] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no.6, pp. 854–864, Dec. 2016.

[9] Y. Yang, K. Wang, G. Zhang, X. Chen, X. Luo, and M.-T. Zhou, "MEETS: Maximal energy efficient task scheduling in homogeneous fog networks," *IEEE Internet Things J.*, vol. 5, pp. 4076–4087, Oct. 2018.

[10] Y. Yang, "Multi-tier computing networks for intelligent IoT," *Nature Electronics*, vol. 2, pp. 4–5, Jan. 2019.

[11] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no.1, pp. 5–36, Jan. 2022.

[12] Y. Xiao and M. Krunz, "Dynamic network slicing for scalable fog computing systems with energy harvesting," *IEEE J. Sel. Areas Commun.*, vol. 36, pp. 2640–2654, Dec. 2018.

[13] Y. Xiao and M. Krunz, "Distributed optimization for energy-efficient fog computing in the tactile internet," *IEEE J. Sel. Areas Commun.*, vol. 36, pp. 2390–2400, Nov. 2018.

[14] B. V. Philip, T. Alpcan, J. Jin, and M. Palaniswami, "Distributed real-time iot for autonomous vehicles," *IEEE Transactions on Industrial Informatics*, vol. 15, no.2, pp. 1131–1140, Feb. 2020.

[15] M. Afrin, J. Jin, A. Rahman, A. Rahman, J. Wan, and E. Hossain, "Resource allocation and service provisioning in multi-agent cloud robotics: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no.2, pp. 842–870, Second Quarter 2021.

[16] K. Wang, Y. Tan, Z. Shao, S. Ci, and Y. Yang, "Learning-based task offloading for delay-sensitive applications in dynamic fog networks," *IEEE Trans. Veh. Tech.*, vol. 68, no.11, pp. 11399–11403, Nov. 2019.

[17] K. Wang, Y. Zhou, J. Li, S. L, W. Chen, and L. Hanzo, "Energy-efficient task offloading in massive MIMO-aided multi-pair fog-computing networks," *IEEE Trans. Commun.*, vol. 69, no.4, pp. 2123–2137, Apr. 2021.

[18] K. Wang, Y. Zhou, Q. Wu, W. Chen, and Y. Yang, "Task offloading in hybrid intelligent reflecting surface and massive MIMO relay networks," *IEEE Trans. Wireless Commun.*, vol. 21, no.6, pp. 3648–3663, Jun. 2022.

[19] Y. Yang, Z. Liu, X. Yang, K. Wang, X. Hong, and X. Ge, "POMT: Paired offloading of multiple tasks in heterogeneous fog networks," *IEEE Internet Things J.*, vol. 6, pp. 8658–8669, Oct. 2019.

[20] Z. Liu, Y. Yang, K. Wang, Z. Shao, and J. Zhang, "POST: Parallel offloading of splittable tasks in heterogeneous fog networks," *IEEE Internet Things J.*, vol. 7, pp. 3170–3182, Apr. 2020.

[21] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, pp. 742–758, Oct. 2014.

[22] O. Y. Bursalioglu, G. Caire, R. K. Mungara, H. C. Papadopoulos, and C. Wang, "Fog massive MIMO: A user-centric seamless hot-spot architecture," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 559–574, Jan. 2019.

[23] D. Chen, "Low complexity power control with decentralized fog computing for distributed massive MIMO," in *Proc. IEEE WCNC*, (Barcelona, Spain), pp. 1–6, Apr. 2018.

[24] R. K. Mungara, G. Caire, O. Y. Bursalioglu, C. Wang, and H. C. Papadopoulos, "Fog massive MIMO with on-the-fly pilot contamination control," in *Proc. IEEE ISIT*, (Vail, CO, USA), pp. 1–5, Jun. 2018.

[25] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, and H. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, 6, pp. 64–71, Jun. 2017.

[26] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no.2, pp. 1100–1114, Feb. 2015.

[27] T. Bai, C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Resource allocation for intelligent reflecting surface aided wireless powered mobile edge computing in OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 20, pp. 5389–5407, Aug. 2021.

[28] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 5394–5409, Nov. 2019.

[29] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE J. Sel. Areas Commun.*, vol. 38, pp. 1809–1822, Aug. 2020.

[30] M. D. Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, pp. 2450–2525, Nov. 2020.

[31] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, pp. 106–112, Jan. 2020.

[32] T. Bai, C. Pan, C. Han, and L. Hanzo, "Reconfigurable intelligent surface aided mobile edge computing," *IEEE Wireless Communications*, 2021.

[33] X. Hu, C. Masouros, and K. K. Wong, "Removing channel estimation by location-only based deep learning for RIS aided mobile edge computing," in *Proc. of the IEEE ICC 2021*, (Montreal, Canada), pp. 1–6, Jun. 2021.

[34] Q. Wu, X. Zhou, and R. Schober, "IRS-assisted wireless powered NOMA: Do we really need different phase shifts in DL and UL?," *IEEE Wireless Commun. Lett.*, vol. 10, pp. 1493–1497, Jul. 2021.

[35] Z. Chu, P. Xiao, M. Shojafar, D. Mi, J. Mao, and W. Hao, "Intelligent reflecting surface assisted mobile edge computing for internet of things," *IEEE Wireless Commun. Lett.*, vol. 10, pp. 619–623, Mar. 2021.

[36] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple access downlink transmissions," *IEEE Trans. Veh. Tech.*, vol. 65, pp. 6010–6023, Aug. 2016.

[37] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Nonorthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, pp. 74–81, Sept. 2016.

[38] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge University Press, 2005.

[39] M. Mohseni, R. Zhang, and J. M. Cioffi, "Optimized transmission for fading multiple-access and broadcast channels with multiple antennas," *IEEE J. Sel. Areas Commun.*, vol. 24, no.8, pp. 1627–1639, Aug. 2006.

[40] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An optimization perspective of the superiority of NOMA compared to conventional OMA," *IEEE Trans. Sig. Proc.*, vol. 65, pp. 5191–5202, Oct. 2017.

[41] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource allocation in NOMA based fog radio access networks," *IEEE Wireless Commun.*, vol. 25, pp. 110–115, Jun. 2018.

[42] X. Wen, H. Zhang, H. Zhang, and F. Fang, "Interference pricing resource allocation and user-subchannel matching for NOMA hierarchy fog networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, pp. 467–479, Jun. 2019.

[43] K. Wang, Y. Zhou, Z. Liu, Z. Shao, X. Luo, and Y. Yang, "Online task scheduling and resource allocation for intelligent NOMA-based industrial internet of things," *IEEE J. Select. Areas Commun*, vol. 38, no.5, pp. 803–815, May 2020.

[44] C. Dong, Y. Shen, Y. Qu, K. Wang, J. Zheng, Q. Wu, and F. Wu, "UAVs as an intelligent service: Boosting edge intelligence for air-ground integrated networks," *IEEE Network*, vol. 35, pp. 167–175, Jul. 2021.

[45] N. Cheng, W. Xu, W. Shi, Y. Zhou, N. Lu, H. Zhou, and X. Shen, "Air-ground integrated mobile edge networks: Architecture, challenges, and opportunities," *IEEE Commun. Mag.*, vol. 56, pp. 26–32, Aug. 2018.

[46] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no.4, pp. 2322–2358, Aug. 2017.

[47] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no.4, pp. 2595–2621, Jun. 2018.

[48] F. Tang, B. Mao, Z. M. Fadlullah, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "On removing routing protocol from future wireless networks: A real-time deep learning approach for intelligent traffic control," *IEEE Wireless Communications*, vol. 25, pp. 154–160, Feb. 2018.

[49] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no.8, pp. 84–90, Aug. 2019.

[50] C. Jiang, N. Ge, and L. Kuang, "AI-enabled next-generation communication networks: Intelligent agent and AI router," *IEEE Wireless Commun.*, vol. 27, no.6, pp. 129–133, Dec. 2020.

[51] D. C. Nguyen, P. Cheng, M. Ding, D. Lopez-Perez, P. N. Pathirana, J. Li, A. Seneviratne, Y. Li, and H. V. Poor, "Enabling AI in future wireless networks: A data life cycle perspective," *IEEE Commun. Surveys Tuts.*, vol. 23, no.1, pp. 553–595, First Quarter 2021.

[52] J. G. A. et al., "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no.6, pp. 1065–1082, Jun. 2014.

[53] L. Pu, X. Chen, J. Xu, and X. Fu, "D2D fogging: An energy-efficient and incentive-aware task offloading framework via network-assisted D2D collaboration," *IEEE J. Sel. Areas Commun.*, vol. 34, no.12, pp. 3887–3901, Dec. 2016.

[54] X. Chen, L. Pu, L. Gao, W. Wu, and D. Wu, "Exploiting massive D2D collaboration for energy-efficient mobile edge computing," *IEEE Wireless Commun.*, vol. 24, pp. 64–71, Aug. 2017.

[55] Y. Yang, S. Zhao, W. Zhang, Y. Chen, X. Luo, and J. Wang, "DEBTS: Delay energy balanced task scheduling in homogeneous fog networks," *IEEE Internet Things J.*, pp. 1–1, 2018.

[56] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, "An overview of MIMO communicatios-a key to Gigabit wireless," in *Proceedings of the IEEE.*, vol. 92, no.2, pp. 198–218, Feb. 2002.

[57] K. Wang, W. Chen, J. Li, and B. Vucetic, "Green MU-MIMO/SIMO switching for heterogeneous delay-aware services with constellation optimization," *IEEE Trans. Commun.*, vol. 64, pp. 1984–1995, May 2016.

[58] K. Wang and W. Chen, "Energy-efficient communications in MIMO systems based on adaptive packets and congestion control with delay constraints," *IEEE Trans. Wireless Commun.*, vol. 14, no.4, pp. 2169–2179, Apr. 2015.

[59] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 3590–3600, Nov. 2010.

[60] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Multiple-antenna techniques in LTE-advanced," *IEEE Signal Process. Mag.*, vol. 30, pp. 40–60, Oct. 2013.

[61] K. R. Liu, *Cooperative communications and networking*. Cambridge, U.K.: Cambridge Univ. Press., 2009.

[62] G. Amarasuriya, "Sum rate analysis for multi-user massive MIMO relay networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, pp. 1–7, Dec. 2015.

[63] N. Yang, M. Elkashlan, P. L. Yeoh, and J. Yuan, "Multiuser MIMO relay networks in nakagami-m fading channels," *IEEE Trans. Commun.*, vol. 60, no.11, pp. 3298–3310, Nov. 2012.

[64] F. Zhou, C. You, and R. Zhang, "Delay-optimal scheduling for IRS-aided mobile edge computing," *IEEE Wireless Communications Letters*, vol. 10, pp. 740–744, Apr. 2021.

[65] T. Bai, C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Resource allocation for intelligent reflecting surface aided wireless powered mobile edge computing in OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 20, pp. 5389–5407, Aug. 2021.

[66] K. Wang, Z. Ding, D. K. C. So, and G. K. Karagiannidis, "Stackelberg game of energy consumption and latency in mec systems with NOMA," *IEEE Trans. on Commun*, vol. 69, no.4, pp. 2191–2206, Apr. 2021.

[67] C. Zhao, Y. Cai, M. Zhao, and Q. Shi, "Joint hybrid beamforming and offloading for mmwave mobile edge computing systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, pp. 1–6, June 2019.

[68] C. Zhao, Y. Cai, A. Liu, M. Zhao, and L. Hanzo, "Mobile edge computing meets mmWave communications: Joint beamforming and resource allocation for system delay minimization," *IEEE Trans. Wireless Commun.*, vol. 19, no.4, pp. 2382–2396, Apr. 2020.

[69] X. Yu, F. Xu, J. Cai, X. Y. Dang, and K. Wang, "Computation efficiency optimization for millimeter-wave mobile edge computing networks with noma," *IEEE Trans. Mobile Computing*, vol. 1, pp. 1–1, Apr. 2022.

[70] S. Yu, X. Gong, Q. Shi, X. Wang, and X. Chen, "EC-SAGINs: Edge computing-enhanced space-air-ground integrated networks for internet of vehicles," *IEEE Internet Things J.*, pp. 1–13, 2021.

[71] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang, and X. S. Shen, "Software defined space-air-ground integrated vehicular networks: Challenges and solutions," *IEEE Commun. Mag.*, vol. 55, pp. 101–109, Jul. 2017.

[72] Z. Zhou, J. Feng, L. Tan, Y. He, and J. Gong, "An air-ground integration approach for mobile edge computing in IoT," *IEEE Commun. Mag.*, vol. 56, pp. 40–47, Aug. 2018.

[73] N. Kato, Z. M. Fadlullah, F. Tang, B. Mao, S. Tani, A. Okamura, and J. Liu, "Optimizing space-air-ground integrated networks by artificial intelligence," *IEEE Wireless Communications*, vol. 26, pp. 140–147, Aug. 2019.

[74] N. Cheng, F. Lyu, W. Quan, C. Zhou, H. He, W. Shi, and X. Shen, "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no.5, pp. 1117–1129, May 2019.

[75] B. Shang and L. Liu, "Mobile-edge computing in the sky: Energy optimization for aircground integrated networks," *IEEE Internet Things J.*, vol. 7, no.8, pp. 7443–7456, Aug. 2020.

[76] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," Apr. 2017. [Online]. Available: https://arxiv.org/pdf/1602.05629.pdf.

[77] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no.8, pp. 1655–1674, Aug. 2019.

[78] L. Huang, X. Feng, A. Feng, Y. Huang, and L. P. Qian, "Distributed deep learning-based offloading for mobile edge computing networks," *Mobile Networks and Applications*, pp. 1–8, Nov. 2018. doi: 10.1007/s11036-018-1177-x.

[79] G. Qu and H. Wu, "DMRO: A deep meta reinforcement learning-based task offloading framework for edge-cloud computing," Aug. 2020. [Online]. Available: https://arxiv.org/abs/2008.09930.

[80] L. Huang, S. Bi, and Y. J. A. Zhang, "Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 19, pp. 2581–2593, Nov. 2020.

[81] J. Wang, L. Zhao, J. Liu, and N. Kato, "Smart resource allocation for mobile edge computing: A deep reinforcement learning approach," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, pp. 1529–1541, Sep. 2021.

[82] S. Yu, X. Chen, Z. Zhou, X. Gong, and D. Wu, "When deep reinforcement learning meets federated learning: Intelligent multitimescale resource management for multiaccess edge computing in 5G ultradense network," *IEEE Internet Things J.*, vol. 8, no.4, pp. 2238–2251, Feb. 2021.

[83] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, pp. 375–388, Jan. 2020.

[84] A. M. Elbir, A. Papazafeiropoulos, P. Kourtessis, and S. Chatzinotas, "Deep channel learning for large intelligent surfaces aided mm-wave massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 9, no.9, pp. 1447–1451, Sep. 2020.

[85] M. Chiang, S. Ha, F. Risso, T. Zhang, and I. Chih-Lin, "Clarifying fog computing and networking: 10 questions and answers," *IEEE Commun. Mag.*, vol. 55, no.4, pp. 18–20, Apr. 2017.

[86] K. Wang, W. Chen, J. Li, Y. Yang, and L. Hanzo, "Joint task offloading and caching for massive MIMO-aided multi-tier computing networks," *IEEE Trans. Commun.*, vol. 70, pp. 1820–1833, Mar. 2022.

[87] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, pp. 485–500, Apr. 2016.

[88] W. Zhao and S. Wang, "Resource allocation for device-to-device communication underlaying cellular networks: An alternating optimization method," *IEEE Commun. Lett.*, vol. 19, pp. 1398–1401, Aug. 2015.

[89] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency minimization for intelligent reflecting surface aided mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, no.11, pp. 2666–2682, Nov. 2020.

[90] X. Ren, D. Li, Y. Xi, and H. Shao, "Distributed global optimization for a class of nonconvex optimization with coupled constraints," *IEEE Transactions on Automatic Control*, 2021.

[91] T. Tatarenko and B. Touri, "Non-convex distributed optimization," *IEEE Transactions on Automatic Control*, vol. 62, no.8, pp. 3744–3757, Aug. 2017.

[92] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Trans. Sig. Proc.*, vol. 66, no.11, pp. 2834–2848, Jun. 2018.

[93] I. V. Sergienko, L. F. Hulianytskyi, and S. I. Sirenko, "Classification of applied methods of combinatorial optimization," *Cybern. Syst. Analysis*, vol. 45, no.5, no. 5, pp. 732–741, 2009.

[94] L. F. Hulyanytskyi and S. I. Sirenko, "ACO-H metaheuristic combinatorial optimization method," *J. Autom. Inform. Sci.*, vol. 42, no.7, no. 7, pp. 30–42, 2010.

[95] T. N. Barbolina, "Solution of mixed combinatorial optimization problems on arrangements by the method of construction of lexicographic equivalence," *Cybern. Syst. Analysis*, vol. 49, no.6, no. 6, pp. 922–931, 2013.

[96] O. A. Yemets, T. N. Barbolina, and O. A. Chernenko, "Soving optimization problems with linear-fractional objective functions and additional constraints on arrangements," *Cybern. Syst. Analysis*, vol. 42, no.5, no. 5, pp. 680–685, 2006.

[97] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no.1, pp. 269–283, Jan. 2021.

[98] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Networking*, vol. 24, pp. 2795–2808, Oct. 2016.

[99] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading decision and resource allocation for multi-user multi-task mobile cloud," in *Proc. of the IEEE ICC*, (Kuala Lumpur, Malaysia), pp. 1–6, May 2016.

[100] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *IEEE Sign. Proc. Lett*, vol. 25, no.12, pp. 1875–1879, Dec. 2018.

[101] J. Zheng, Y. Cai, Y. Wu, and X. Shen, "Dynamic computation offloading for mobile cloud computing: A stochastic game-theoretic approach," *IEEE Trans. Mobile Comput.*, vol. 18, no.4, pp. 771–786, Apr. 2019.

[102] J. Ren, G. Yu, Y. He, and G. Y. Li, "Collaborative cloud and edge computing for latency minimization," *IEEE Trans. Veh. Tech.*, vol. 68, no.5, pp. 5031–5044, May 2019.

[103] M. Sheng, Y. Dai, J. Liu, N. Cheng, X. Shen, and Q. Yang, "Delay-aware computation offloading in NOMA MEC under differentiated uploading delay," *IEEE Trans. Wireless Commun.*, vol. 19, no.4, pp. 2813–2826, Apr. 2020.

[104] Z. Yang, Y. Liu, Y. Chen, and N. Al-Dhahir, "Cache-aided NOMA mobile edge computing: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no.10, pp. 6899–6915, Oct. 2020.

[105] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT press, 2016.

[106] X. Cao, J. Zhang, and H. V. Poor, "Online stochastic optimization with time-varying distributions," *IEEE Transactions on Automatic Control*, vol. 66, no.4, pp. 1840–1847, Apr. 2021.

[107] Z. Liu, X. Yang, Y. Yang, K. Wang, and G. Mao, "DATS: Dispersive stable task scheduling in heterogeneous fog networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3423–3436, 2019.

[108] M. Mukherjee, L. Shu, and D. Wang, "Survey of fog computing: Fundamental, network applications, and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1826–1857, 2018.

[109] S. Wang, R. Urgaonkar, M. Zafer, T. He, K. Chan, and K. K. Leung, "Dynamic service migration in mobile edge computing based on markov decision process," *IEEE/ACM Trans. Networking*, vol. 27, no.3, pp. 1272–1288, Jun. 2019.

[110] I. A. Ridhawi, M. Aloqaily, Y. Kotb, Y. Jararweh, and T. Baker, "A profitable and energy-efficient cooperative fog solution for IoT services," *IEEE Transactions on Industrial Informatics*, vol. 16, pp. 3578–3586, May 2020.

[111] X. Xiong, K. Zheng, L. Lei, and L. Hou, "Resource allocation based on deep reinforcement learning in IoT edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, no.6, pp. 1133–1146, Jun. 2020.

[112] X. Zhang, J. Zhang, Z. Liu, Q. Cui, X. Tao, and S. Wang, "MDP-based task offloading for vehicular edge computing under certain and uncertain transition probabilities," *IEEE Trans. Veh. Tech.*, vol. 69, no.3, pp. 3296–3309, Mar. 2020.

[113] S. N. Shirazi, A. Gouglidis, A. Farshad, and D. Hutchison, "The extended cloud: Review and analysis of mobile edge computing and fog from a security and resilience perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no.11, pp. 2586–2595, Nov. 2017.