This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

CHINA COMMUNICATIONS

BEYOND 5G NETWORKS: INTEGRATION OF COMMUNICATION, COMPUTING, CACHING, AND CONTROL

Musbahu Mohammed Adam[1], Liqiang Zhao [1,*], Kezhi Wang[2], and Zhu Han[3]

[1] State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China
[2] Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK
[3] Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 02447 South Korea

Abstract: In recent years, the exponential proliferation of smart devices with their intelligent applications poses severe challenges on conventional cellular networks. Such challenges can be potentially overcome by integrating communication, computing, caching, and control (i4C) technologies. In this survey, we first give a snapshot of different aspects of the i4C, comprising background, motivation, leading technological enablers, potential applications, and use cases. Next, we describe different models of communication, computing, caching, and control (4C) to lay the foundation of the integration approach. We review current state-of-the-art research efforts related to the i4C, focusing on recent trends of both conventional and artificial intelligence (AI)-based integration approaches. We also highlight the need for intelligence in resources integration. Then, we discuss the integration of sensing and communication (ISAC) and classify the integration approaches into various classes. Finally, we propose open challenges and present future research directions for beyond 5G networks, such as 6G.

Keywords: 4C; 6G; integration of communication, computing, caching, and control; i4C; multi-access edge computing (MEC)

Received: TBD
Revised: TBD
Editor: TBD

I. INTRODUCTION

Currently, there exist many research efforts in academia and industry that have been devoted to addressing the longstanding issues of communication, computing, caching, and control (4C) functionalities. However, a considerable number of these efforts focused on improving the performance of these underlying functionalities separately, which in turn leads to their unavoidable shortcomings. For instance, in the communication domain, the popular Shannon capacity limit is on the verge of being approached with the existing long-term evolution (LTE) techniques [1–3]. In the area of computing, the Moore's law is fast approaching its impending limit based on silicon chips technology. In the caching/storage domain, the rapid progress in magneto-optical and optical disks for storage may not accommodate the increasingly growing big data demands [1]. Likewise, the performance of control could be limited by multiple factors, including heterogeneous users' demands, wireless fading channels, and insufficient computing power. In particular, wireless networked control systems share information among sensors, actuators, and plants using wireless networks, characterized by deep fade and susceptible to signal power loss. Such channel impairments render the control functionality/algorithm sub-optimal [4]. Besides, control algorithm relies on extensive computations to run some networked control systems rapidly; thus, insufficient/weak processing units degrade the control performance.

Due to these inherent limitations, further signifi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

cant performance improvement in terms of communication, computing, or caching capabilities becomes more challenging for engineers and researchers in research and development sectors [1]. In other words, optimizing any one of the 4C functionalities/resources will hardly maximize the performance of a communication network [5]. Hence, relying on a single 4C functionality alone will no longer sustain the requirements of emerging intelligent applications and services. Nonetheless, the great advances in the individual domains of 4C triggered some promising steps toward proposing hybrid functionalities [6], leading to the revolutionary changes that warrant the respective functionalities of 4C to encroach on one another's territory. Here is why one can hardly place a clear boundary among communication, storage (cache), computing [1], and control domains nowadays.

Moreover, the individual functionalities of 4C have great potential to complement and reinforce one another. For example, edge caching technique can minimize traffic redundancy, avoid duplicate transmission, and reduce bandwidth consumption in communications [5]. On top of that, the control algorithm is essential for controlling, coordinating, and optimizing the other integrants of 4C. The promising gains derived from the capabilities of 4C accelerate the progress toward integrating them in future networks. Achieving this striking breakthrough implies a paradigm shift toward the information transmission, processing, storage, and intelligent decision-making networks that support the new-technologies-new-services trends. Indeed, the integration of communication, computing, caching, and control (i.e., the i4C) will provide massive support for the fifth-generation (5G), sixth-generation (6G), and beyond networks, enabling key network elements, functionalities, and heterogeneous services.

1.1 Integration for 5G, 6G, and Beyond

Against prior network generations, the 5G network is emerging with a much more complex mission, i.e., supporting the dramatic evolution of information and communications technology (ICT) and the Internet. Hence, 5G systems support not only communication functionality but also the other three parts of the 4C functionalities. These functionalities play pivotal roles in enabling a variety of services in 5G, e.g., massive machine type communication (mMTC), enhanced mobile broadband (eMBB), and ultra-reliable and low latency communication (URLLC) [7, 8]. Today, with the advent of 5G networks, we witness a great boom in intelligent applications and use cases. Such applications keep emerging with unusual demands for communication, computing [8], and caching resources. So far, the 5G's achievement in satisfying the stringent requirements of the applications is insufficient.

Furthermore, the evolution of 5G prompts the notion of beyond 5G networks, such as 6G and beyond, for superior performance. In 6G networks, novel disruptive wireless technologies and futuristic network architectures will be put into perspective. 6G is further envisaged to ultimately attain future generation connectivity, driven and motivated by the transformation from "*connected everything*" to "*connected intelligence,*" hence facilitating "*human-thing intelligence*" interconnectivity [9].

Therefore, compared with 5G network, 6G is expected to surface with larger dimensions, higher complexity, dynamicity, and heterogeneity features. These issues require a novel, agile, flexible, adaptive, and intelligent architecture. Featured with intelligent recognition, high learning, predicting, and powerful reasoning and decision-making abilities, *artificial intelligence (AI)* can allow the 6G network architecture to learn, intelligently decide, and adjust itself in order to enable diversified services without requiring human support [10]. To this end, Letaief et al. [11] applied AI techniques to realize network intelligentization, closed-loop optimization, and intelligent wireless communications for 6G networks.

Moreover, AI will serve as a powerful assistant for the communication, computing, and caching functionalities in edge computing. This promising novel paradigm is termed *intelligent edge*. With the AI techniques, intelligent edge provides optimal edge computing solutions, such as resource allocation optimization [12].

The idea of softwarization alone will no longer be sufficient for the 6G mobile networks due to the growing complexity and heterogeneity in mobile wireless communication networks. To be more specific, network elements should support different capabilities, comprising AI control, communication, computing, content caching, and even wireless power transfer, for supporting mobile AI-based applications [11]. In other

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

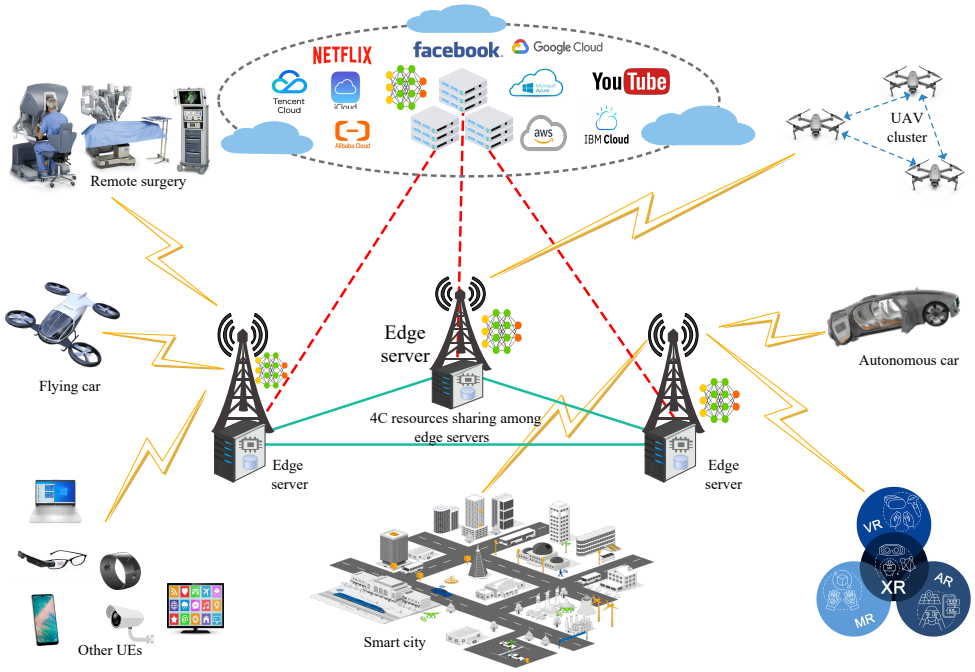

Figure 1. *A typical scenario for the i4C in 5G, 6G, and beyond.*

words, converging intelligent control, sensing, communication, computing, and caching functions will be a core driver behind the emerging 6G networks. The recent influx of the AI frontiers (e.g., deep learning (DL), federated learning (FL), machine leraning (ML), and deep reinforcement learning (DRL) techniques) at the network edges implies the urgent need for the intelligent decision-making techniques that can efficiently interact the conflicting functionalities in future networks. In fact, 6G networks will rely on the AI frontiers to realize intelligent optimization of 4C. Beyond that, integrating communication, computing, and caching with AI-based/intelligent control could be one of the biggest revolutionary trends that come to stay in 6G and beyond mobile networks for disruptive applications/innovations.

In short, 6G networks will unleash the maximum potential of communication with computing and control at the proximity of myriad mission-critical applications [11], such as extended reality (XR), Tactile Internet, autonomous vehicles, Internet of Vehicles (IoV), Internet of Everything (IoE), Internet of Intelligent Things (IoIT), flying vehicles, and space-air-ground integrated networks. The i4C will indeed allow the 6G mobile networks to make appropriate decisions on the user equipment (UE) applications' tasks. The control will ensure that only tasks to be completely and optimally executed, given the available resources at the moment, are granted resources. Hence, the convergence of 4C will bring unprecedented solutions to multitudinous intelligent applications that emerge with stringent requirements for massive connectivity, ultra-high reliability, ultra-low latency, high mobility, energy-saving, and so on. Above all, the i4C will drive wireless networks to reach newer and greater heights of performance. Fig. 1 portrays a typical scenario where the integrated 4C resources embedded in distributed collaborative network edge servers are shared for different quality of service (QoS) requirements. Therefore, moving toward integrating 4C (i.e., i4C) should be the focal point for the evolution of 6G and beyond mobile networks.

1.2 Motivation and Contributions

In recent years, we witness many promising gains brought by the evolution of information technology (IT) in wireless network environments. The appreciable advances in IT and communications have been rev-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

Table 1. *Comparison of existing surveys on the integration of resources*

Related Surveys	Themes	Key Contributions	Limitations
[7]	Communication and Computation Resource Allocation	● Surveyed communication and computation models in MEC. ●Reviewed communications and computation resources allocation.	●Omitted caching and control and ignored their models. ● Omitted recent efforts on AI and edge intelligence.
[8]	Integrating MEC into 5G Technologies	● Focused on fusing MEC and 5G technologies. ● Applications of ML in MEC, comprising 4C optimization, big data, etc.	● Discussed 4C optimization as an application of ML in MEC. ● Omitted 4C models.
[13–15]	Converging Communications, Computing, and Caching in Mobile Edge Networks	● Reviewed issues of converging/integrating communication, computing, and caching. ● Focused on definition, architecture/frameworks, enablers, metrics, IoT, and challenges.	● Did not cover recent efforts on edge intelligence and AI solutions for i4C. ● Omitted control. ● Ignored 4C models.
[16]	Integrating Communication, Computing, and Control	● Surveyed the integration of communication, computing, and control. ● Focused on real-time computing, networked control, real-time networking, wireless sensor networks, and autonomous vehicles.	● Ignored efforts on network edges, edge intelligence, and the AI roles in i4C. ● Did not cover caching. ● Omitted 4C models.
This Survey	i4C	● Reviewed the i4C with emphasis on AI-based and conventional integration approaches. ● Surveyed the integration aspects, including motivations, key enablers, applications and use cases. ● Discussed 4C models, challenges, and future direction.	

olutionizing conventional networks in terms of structures and operations, driving new capabilities by leveraging synergies among different functionalities of 4C. Considering significant benefits brought by the individual functionalities of 4C, integrating them into a single system/network becomes necessary for users' satisfaction and networks' performance requirements. Converging 4C may lead to realizing the optimal solutions that will fulfill diverse QoS requirements of futuristic intelligent applications and use cases in the coming decades. Suffice to say, the i4C becomes a natural trend that has pivotal roles to play in 5G, 6G, and beyond networks.

Today, different aspects of 4C receive attention from academia and industry. In fact, several research efforts investigated the trends and issues of converging communication, computing, and caching and that of communication, computing, and control. However, a few survey articles, including [7], [8], and [13–16], focused on these concepts. Table 1 summarizes the contributions of these surveys. Specifically, Mao et al. [7] reviewed efforts on mobile edge computing (MEC), focusing on joint computation and communications resource allocation in MEC and their models. Pham et al. [8] focused on integrating MEC with potential 5G technologies and discussed applications of machine learning in MEC, comprising 4C optimization, crowdsensing, big data, and privacy and security. The survey in [13] reviewed recent works on mobile network edges, exploring issues of communication, computation, and caching techniques, and discussed challenges and applications of edge networks. Wang et al. [14] reviewed some efforts intended to integrate communication, computing, and caching. The central theme of the survey focused on key aspects of the integration, i.e., motivations, enabling technologies, performance metrics, frameworks, and challenges. In [15], the authors reviewed the trends of technology in communication, caching, and computing resources, analyzing the interactions among the resources while collecting, storing, indexing, and processing IoT data. They also described the convergence in devices, sensors, and gateways. [16] focused on integrating communication, computing, and control and its potential applications.

These surveys contributed greatly to optimizing the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

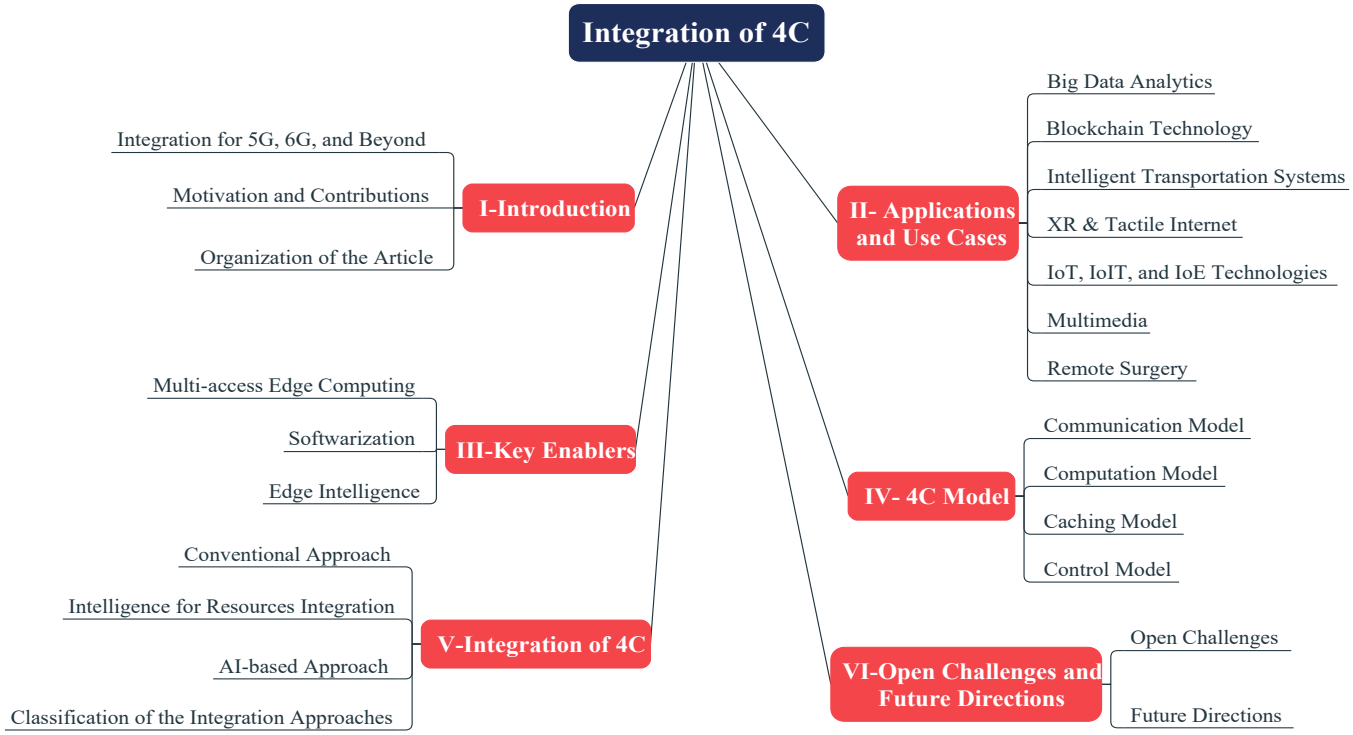

Figure 2. *The roadmap of the survey.*

networks' performance and improving the users' experience. Nevertheless, neither integrating communication, computing, and caching nor integrating communication, computing, and control is sufficient. Of course, the roles of control in resources integration cannot be overstated. For example, in radio access network (RAN) slicing and distributed resource allocation, a control scheme has to be put in place to guide the allocation and deallocation of the competitive network resources. Besides, to satisfy diverse QoS needs, 5G New Radio (NR) is designed to be flexible; the 6G network is expected to be much more flexible and complex. The growing flexibility in these networks implies more control parameters necessitating essential changes in wireless network operations [17]. Likewise, caching saves network bandwidth and avoids the transmission of duplicate content. Therefore, the need for i4C arises to shape up the visions of future generation networks.

To the best of our knowledge, there is no existing survey that devoted itself to converging 4C. Hence, this paper aims to present a firsthand tutorial on the convergence of 4C against the aforementioned surveys. The survey differs from the previously mentioned ones with the following contributions. To begin with, the paper discusses different aspects of the integration, exploring its background, motivations, leading enabling technologies, and potential benefits and use cases. Another point worth noticing is that the existing surveys omitted the 4C models, which serve as the backbone of the design and implementation of an integrated 4C network/system. Considering the roles of AI in 6G networks and the increasingly growing complexity of wireless networks, the paper pays much attention to the recent trends of i4C based on the AI techniques. Thus, the paper comprehensively reviews the i4C, focusing on conventional and AI-based approaches. It also classifies various approaches of the integration and discusses the integration of sensing and communication (ISAC). Then, it considers several open challenges and provides future directions. The roadmap of this survey is shown in Fig. 2.

1.3 Organization of the Article

The rest of this paper is organized as follows: Section II brings a snapshot of potential applications and use cases. Section III presents the main enablers for i4C in future mobile networks. Section IV discusses vari-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

ous models of 4C, which lays the foundation for their integration. Section V reviews many cutting-edge research efforts on the i4C with a focus on both AI-based and conventional optimization/integration approaches. It also discusses the convengence of communication and sensing and classifies different approaches of integrating resources. Section VI focuses on open challenges and explores future research directions. Finally, Section VII concludes the survey.

II. POTENTIAL APPLICATIONS AND USE CASES

Numerous drivers motivate the convergence of 4C functionalities. One of the key driving forces behind the i4C is the explosion of wide-ranging intelligent applications and use cases, including big data, autonomous cars, telesurgery, Internet of Things (IoT), XR & Metaverse, Tactile Internet, multimedia, and energy/power systems. Generally, such applications come up with different service requirements, including ultra-low latency, higher throughput, ultra-high reliability, intensive computational capabilities, and vast caching resources. This section highlights the roles of i4C in accommodating the needs of these applications.

2.1 Big Data Analytics

Data analytics undergoes revolution in numerous scientific domains due to the exponential growth of data. This complex and enormous volume of data calls for parallelization at an unprecedented scale because processing it may exceed the capabilities of a single or even a couple of machines [18]. Today, mobile UEs and other IoT devices offload tasks and corresponding data via wireless channels with varying rates, and such data, owing to its diversity, scale, and timeliness, may require real-time analytics and live stream computing and caching. These raise the need for fast, parallel, and distributed processing [19]. To this end, 4C is pushed closer to the devices generating the data, avoiding the bottlenecks resulting from the need of moving data from the storage to the central processing unit (CPU) and its main memory and back. In this way, real-time response, lower latency, and energy saving can be guaranteed [20].

To support big data analytics, characterized by increasingly growing volume, variety, and velocity

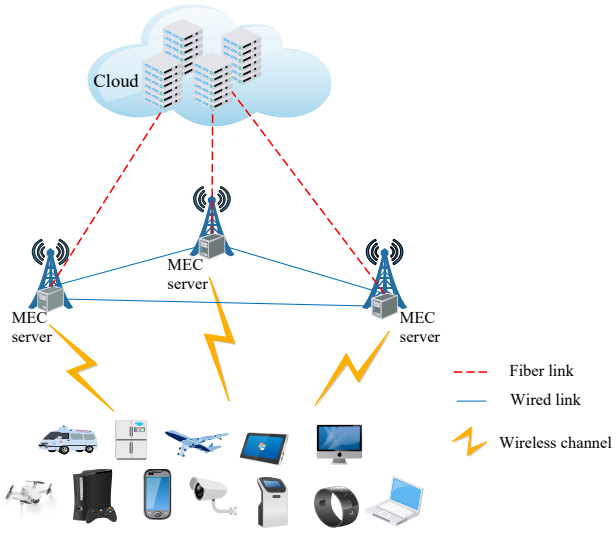

Figure 3. *Collaborative MEC servers for big data processing.*

(3Vs) of data, the big data infrastructure has to couple a scalable data storage with an ultrafast data processing system [21]. Thus, the analytics system should be configurable, flexible, and scalable both vertically and horizontally [22]. Of course, a highly distributed collaborative system of 4C deployed at the network edges will serve significant roles for big data requirements in future networks. Furthermore, the convergence of 4C capabilities at the network edges can allow 6G networks to handle huge volumes of data with high-speed data rate connectivity per UE [23]. Fig. 3 shows a collaborative system of MEC servers hosting 4C resources for processing big data. In a word, the i4C is a potentially promising approach to cope with the big data challenges in 5G and 6G networks.

2.2 Blockchain

Besides its bright potential for implementing key services in 5G, 6G, and beyond networks, blockchain technology deserves recognition for its remarkable achievement in cryptocurrencies, such as Bitcoin, Cardano, Ethereum, and other trending metaverse applications. However, due to their distributed nature, such cryptocurrencies call for high bandwidth, sensing, computing, and caching capabilities for pledging the ledger integrity. The cryptocurrency protocol handles huge volumes of data transmitted/broadcasted across the participating/playing nodes [24].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

Despite its enormous potential, blockchain faces some critical issues, including scalability and latency, that limit the performance of its applications [24, 25]. To realize a scalable blockchain system with real-time applications, the 4C functionalities can be tightly converged in distributed systems to handle the demands for ubiquitous connectivity, caching, and computing capabilities.

2.3 Intelligent Transportation Systems

Recently, autonomous vehicles begin to surface with the advent of 5G networks. The key performance requirements of such vehicles include ultra-high reliability and lower transmission delay in terms of millisecond scale. However, the dynamic on-vehicle information processing rates and the randomness of wireless communication channels limit the performance of autonomous vehicles. Due to these inherent limitations, the vehicle-to-vehicle (V2V) communication links will unavoidably experience time-varying delays. Using delayed information in designing the control system of autonomous vehicles could jeopardize the stability of the platoon system [26]. This issue requires a robust integrated system of 4C to sustain the stability of the platoon.

In future transportation systems, intelligent infrastructures and intelligent vehicles will converge. By operating the intelligent traffic infrastructures, the whole traffic systems' throughput can be managed efficiently. On the other hand, by equipping intelligent vehicles with seamlessly integrated systems of embedded computing and in-vehicle networks, wireless data exchange can be enabled between vehicle-to-infrastructure and vehicle-to-vehicle. These two capabilities can allow the vehicles to guide drivers or even drive independently (autonomously) by observing and evaluating traffic conditions, planning ahead of their behavior, and actualizing (implementing) the plan using the drive-by-wire functionalities, which include steering, speed and stability controls, braking, and so on [16]. Hence, the i4C becomes necessary for sustainability, ultra-low latency, efficiency, ultra-high reliability, stability, and safety of autonomous vehicles. In 6G and beyond networks, the convergence of 4C will undoubtedly be fully leveraged to realize the visions of fully autonomous vehicles and the IoVs.

2.4 XR & Tactile Internet

XR is an umbrella term referring to the set of immersive technologies, comprising AR, VR, and mixed reality (MR). To transmit higher resolution/frame rate videos, mobile AR/VR applications impose high demands for greater bandwidth with ultra-low latency, ultra-high reliability, and high computing power. Likewise, the Tactile Internet applications, which may extend to healthcare, entertainment, robotics, and autonomous vehicles, require higher communication bandwidth with ultra-low latency. Thanks to the nature of haptic signals and human perception, the required latency for Tactile Internet is 1 ms. Unfortunately, satisfying the stringent requirements of these applications is beyond the capabilities of the existing LTE networks [27]. The 5G and 6G wireless networks will support the tight convergence of 4C at the vicinity of the users' applications. This implies promising solutions for latency-sensitive applications. Therefore, with an integrated system of 4C at the network edges, the requirements of both Tactile Internet and XR applications will be fulfilled.

2.5 IoT & IoE Technologies

One of the intrinsic features of IoT is its potentiality to provide users with in-built intelligence by enabling its devices (aka IoT devices) to connect. Most of the resource-hungry IoT devices interact with one another, reforming the way individuals perceive their environment and get information. Such smart devices can sense and access information from their surrounding environment and accordingly form what is termed sensory swarm. The accessed information may be conveyed to different applications for processing and analysis [28]. The IoT devices share and interpret information based on standardized formats, and by leveraging the essential functionalities of 4C, IoT transforms its devices from traditional to smart. Hence, the i4C will impact the advancement of flexible and efficient IoT in smart cities, providing end-users with diverse smart services thereby enriching energy, entertainment, environment, healthcare, and transportation [29]. Furthermore, the impact of i4C will potentially reach beyond the concept of IoT. Of course, the emerging IoE paradigm, where everything is connected to everything, requires the convergence of 4C; likewise, the concept of IoIT.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

2.6 Multimedia

The current prevalence of UEs results in the rapid growth of the internet traffic [30], which is dominated by video streaming. Today, video streaming accounts for more than 70% of North American downstream traffic at peak time. However, with unstable wireless network conditions, insufficient bandwidth, and billions of viewing devices, the user experiences are inherently deteriorated, sparking a tussle between the increasing video traffic demand and the quality of viewing experiences. In such conditions, adaptive bit rate (ABR) streaming can be used to enhance viewing experiences. Nevertheless, ABR requires tremendous caching and computing resources for pre-transcoding of each video and caching all video chunks [31]. Thus, future mobile networks have to integrate 4C functionalities in the vicinity of the UEs in order to enable efficient multimedia service delivery (in terms of seamless connectivity, lower latency, and high reliability).

2.7 Remote Surgery

In the coming decades, the 6G networks and beyond will be providing improved healthcare services for human beings. Remote surgery/Telesurgery enables a doctor to perform a surgical operation on a patient without being in the same physical location. This requires ultra-high-speed data rates, ultra-low latency, ultra-high reliability, flexibility, high computational power, and the rest, which can hardly be guaranteed by the 5G capabilities. In 6G and beyond, the 4C functionalities will integrate to steer the remote surgery to greater heights. Moreover, many future applications and use cases, among others holographic teleportation, intelligent production, and intelligent life, will undoubtedly require the i4C.

III. KEY ENABLERS FOR INTEGRATING 4C

This section focuses on the key enabling technologies for the i4C in 5G, 6G, and beyond networks. Specifically, Section III-1 gives a snapshot of multi-access edge computing, Section III-2 discusses softwarization, and Section III-3 focuses on edge intelligence. These fundamental enabling technologies have, indeed, proven their capabilities in providing diversified network solutions.

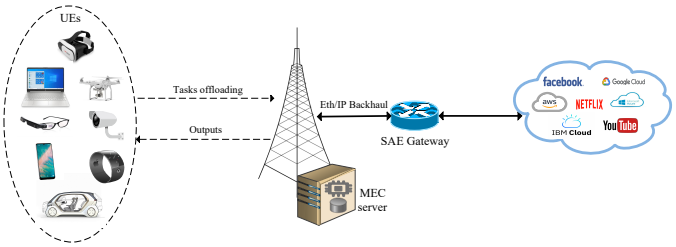

Figure 4. *A typical MEC system.*

3.1 Multi-access Edge Computing (MEC)

MEC emerged to bring the IT and cloud computing capabilities into the RAN domain [7]. Due to the emergence of MEC, 4C is moved to mobile network edges nowadays [7], [19, 32]. With the 4C functionalities at the network edges, data and computational tasks can be wirelessly offloaded, analyzed, computed, and stored near the UEs. In this way, both energy consumption and end-to-end latency will be significantly reduced. However, integrating MEC with a wireless network environment poses some challenges about the control and coordination of joint communication, computing, and caching [19]. Hence, there is a need for optimal decision making on both networks and UEs' computational tasks/data. In the design of an efficient MEC framework, it is essential to flawlessly couple computation offloading control and communication resource management in order to adapt the random variations of wireless channels in terms of frequency, space, and time [7].

On the other hand, the variations in both wireless channels and available computing resources necessitate the need to intelligently control the input data size in UEs for local computing and efficient computation offloading. In doing so, the overall energy consumption for local CPU and transmission will be reduced under a task-deadline constraint [33]. Therefore, joint control of local computing and computation offloading has to be considered for swift execution and caching of the computational tasks. Fig. 4 shows a typical MEC system.

Besides, both computing and cache resources rely on the available communication resources. Specifically, the communication resource is required to offload data and tasks for processing, analyzing, and caching at the network edges [19]. In a 5G mobile edge network, a smart base station (SBS), widely considered a primary 5G infrastructure, will enable the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

integrated 4C services. This requires leveraging the ultimate synergy of the integrated caching, communication, and computing operations by comprehensively considering all essential control factors. In other words, an integrated control scheme is needed to harness the existing synergies between the communication, computing, and caching capabilities for realizing the optimal performance of the 5G network. Thus, in the 5G edge network, a control functionality has to interact the rational agents with conflicting objectives in the SBSs [34]. Accordingly, [35] emerged with a framework that programmably controls and integrates in-network caching, networking, and computing for essential network operations. This further entails the necessity of converging 4C in 5G, 6G, and beyond mobile networks.

3.2 Softwarization

Network functions virtualization (NFV), software-defined networking (SDN), and information-centric networking (ICN) make the leading candidates for softwarization [36], playing significant roles in the i4C. The NFV adopts virtualization techniques to flexibly program the network service functionalities as software instances, i.e., virtual network functions (VNFs), at the network edge servers. The MEC and NFV converge at the network edge to enable the provisioning of computation-oriented services. Thus, various compute-intensive applications will be greatly supported, thereby reducing both operating and capital expenses. Conversely, SDN decouples the control and data planes to improve the network-layer data traffic forwarding and optimizes the network-level resource orchestration; it utilizes a centralized controller in the control plane to receive the network information. The controller, having a global view of the network, makes the network-level decisions for resource allocation, access control policies for UEs, and traffic routing path configuration among the network components, which include network edge servers, access points/BSs, and network switches in the RAN and core network, for enhancing QoS and improving overall resource utilization [37]. Other than splitting the control and data planes, SDN follows the abstraction principles, comprising data traffic forwarding, routing, and configuration as a computing problem. Employing these abstractions results in enabling the network slicing functionalities, and slicing the network leads to enabling the network resource allocation [36].

Hence, the network resource allocation can be effectively managed and optimized by SDN. Adopting SDN to serve as a control module for the integration of resources becomes natural due to its efficiency and effectiveness in managing wireless networks. Nevertheless, the concept of control and resource allocation mechanism of the unified 4C solution is wider than the concept SDN alone; likewise, its diversity goes beyond the reach of any single technology [13]. This implies the need for hybridizing enabling technologies to realize the i4C solutions in future networks.

5G network requires an integrated approach composed of MEC, cloud, and core network. In this paradigm shift, NFV and SDN have challenging roles to play in transforming the way of managing wireless networks [36]. Coupling an NFV and SDN framework with MEC brings centralized network control over communication, computing, and caching resources, thus improving multi-resource orchestration efficiency [37]. In short, integrating NFV and SDN promises flexible network infrastructure, resource management as well as new applications deployment [38].

On the other hand, ICN enables content retrieving for UEs based on identifiers (content identification), not on the basis of physical locations. In 5G, 6G, and beyond, ICN will serve similar roles as NFV and SDN [36]. Note that these promising networking paradigms do not compete, they complement one another instead; they handle various networking issues while benefitting one another. For example, ICN and SDN can be combined to form an SD-ICN framework, hence realizing holistically optimal resource allocation through the logically centralized controller. Moreover, integrating SDN and ICN brings several gains, solving the host-centric networking (TCP/IP) problems and tackling the caching and control issues; see [38].

Today, ICN, NFV, and SDN converge to provide promising 4C solutions. ICN offers a new approach for provisioning services in SDN and NFV; it can also virtualize the network edge functions with some degree of data plane programmability [38]. Due to its ability to allow better migration to cutting-edge technologies through isolation of network parts, SDN-based virtualization excels as a decent approach for converging heterogeneous networks (Het-Nets) with MEC and ICN [39], [40]. It brings several benefits:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

i) it allows flexible management and maintenance of Het-Net framework; ii) with the abstraction and standardization of the data and control planes, networks and applications can be evolved and updated without redesigning the network infrastructure; iii) the control plane can be pushed to the edge/cloud servers rather than a dedicated platform; and iv) both operation and capital expenditures are reduced by utilizing advance software and conventional hardware [40]. In [40], an integrated resources mechanism was built upon the concept of SDN and wireless network virtualization, where MEC and ICN reinforce each other for promoting network efficiency and guaranteeing diverse service requirements.

In 5G and 6G networks, the promising capabilities of ICN, NFV, and SDN will be leveraged to tightly integrate 4C to meet the QoS needs of diverse applications. However, softwarization alone is insufficient for 6G due to the increasing complexity and heterogeneity of wireless networks [11]. This opens a new avenue for integrating communication, computing, and caching with intelligent control in 6G and beyond. Here is where the concept of *edge intelligence* and *intelligent edge* begins.

3.3 Edge Intelligence

One of the essential network entities missed in the 5G mobile networks is edge intelligence, powered by the AI frontiers. Edge intelligence will, in all likelihood, serve the role of a key component in 6G networks to enable new functions, services, and superior performance [41]. Today, edge intelligence is increasingly becoming a center of attention, attracting several research endeavors, due to its promising future and great benefits in 6G and beyond networks.

There are synergistic benefits between the AI techniques and edge computing. For example, edge computing unleashes its scalability and potentials with AI, and the AI technique allows innovations and algorithms for edge computing. Besides, AI extends its applicability to edge computing, and edge computing offers scenarios and platforms for AI. Thus, the AI techniques and edge computing will support and reinforce each other. The prospect of combining edge computing and AI has triggered a solid interest in both academia and industry. The integration of AI and edge computing , which is considered natural and unavoidable, results in the birth of edge intelligence. Actually, edge intelligence goes beyond a mere fusion of the AI techniques and edge computing. The concept of edge intelligence is wide enough and greatly sophisticated, covering several technologies and concepts, which are intertwined together in a mind-boggling way; see [42].

Deng et al. [42] stated that there has not been a formal and globally accepted definition of edge intelligence today. However, the definition is given in some studies. To be specific, Xu et al. [12] defined edge intelligence as a new paradigm of intelligence involving a collection of connected systems and UEs for collecting, analyzing, processing, and caching data near the sources of the data. The goal is to improve the data processing speed and quality and to secure and protect data privacy. Hu et al. [43] described edge intelligence as the paradigm shift involving data collection, transmission, processing, and caching through the use of edge computing with ML techniques and higher networking capabilities.

Contrasted with cloud-based intelligence, in edge intelligence, data is locally analyzed and processed. In edge intelligence, a distributed computing paradigm offers edge inference, edge training, edge caching, and computing services at other edge devices or edge servers for the requirements of a particular edge intelligence application. Thus, edge intelligence brings striking gains by effectively protecting the subscribers' privacy, saving bandwidth, ensuring higher reliability, and lessening response time. Furthermore, by training ML and/or DL models with self-created data, edge intelligence allows subscribers to customize smart applications. In intelligent edge, AI offers strong support for edge computing. The focus of intelligent edge lies in solving edge computing problems with the AI techniques, such as resource allocation optimization [12]. Both intelligent edge and edge intelligence require each other. In fact, the DL services in intelligent edge are likewise a piece of edge intelligence. Hence, in addition to resource utilization, intelligent edge can offer enhanced service throughput for edge intelligence [44].

IV. COMMUNICATION, COMPUTATION, CACHING, AND CONTROL MODELS

A model may represent a theory and plays a crucial role in simplifying the real-life situation analysis. In

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

this section, we explore four different models of 4C to lay the foundation of resource integration for 5G, 6G, and beyond networks. Thus, beneath this section, we discuss the communication model in Section IV-1, the computation model in Section IV-2, the caching model in Section IV-3, and the control model in Section IV-4. As per [7], such models can support mechanisms for abstracting diverse functions and operations into optimization problems and simplifying theoretical analysis.

4.1 Communication Model

In recent decades, stochastic geometry was introduced to serve as a standard tool for modeling and designing wireless networks. A rich set of spatial point processes comprising Poisson Point Process (PPP) and cluster processes have been employed for modeling node locations in various wireless networks, including Het-Nets, cellular networks, and cognitive radio networks. Many research efforts in this area were devoted to addressing interference and wireless channels hostility, such as fading and path loss, to guarantee higher coverage and channel reliability for RAN or distributed D2D networks [45]. Actually, improving the networks' performance in terms of throughput, low latency, and spectral/energy efficiency has been the main emphasis throughout the evolution of mobile communication networks [46]. Hence, several parameters need more attention for efficient wireless networks design and network resources optimization.

4.1.1 Spectral Efficiency (SE)

Various wireless networks air interface techniques, including adaptive modulation and coding (AMC), multiple-input-multiple-output (MIMO) antenna strategies, and frequency domain packet scheduling (FDPS), have improved SE to a great extent. Today, such techniques extend SE near the theoretical Shannon's capacity limit [47]. However, Shannon's theory remains a key design base for the emerging 6G wireless network and offers two main approaches of maximizing network capacity: i) increasing network bandwidth and ii) improving SE. As a key performance indicator (KPI) for the 6G wireless network design and analysis, SE has to be further improved to tackle issues of communication resources, such as multi-dimensional radio and x-haul resources. Thus, in 5G and 6G networks, SE will be enhanced to 3× that of 4G and 5-10× that of 5G, respectively [48]. As per [25], 30 *bps/Hz and 15 bps/Hz* in the downlink and uplink, respectively, mark the minimum requirements for peak SE in 5G. Realizing this will enable efficient computation offloading, which serves a significant role in the i4C. Based on [49], we can express the achievable SE for offloading tasks in an uplink direction in (1) as

$$\ell_u^k = \log_2\left(1 + \frac{p_u^k G_u^k}{\sigma^2}\right), \tag{1}$$

where p_u^k denotes the transmission power from UE to an SBS k and the G_u^k represents the corresponding channel gain between UE u and SBS k.

4.1.2 Interference

Now that the state-of-the-art wireless technologies operate nearer the Shannon capacity bound, a limited capacity gain can be extracted with current cell structures and frequency allocation techniques. The 5G radio access technology (RAT) promises to utilize a three-dimensional capacity model (i.e., bits per second × Hertz × cells per square kilometer), implying that the so-called capacity gain can be achieved with an increase in a number of cells per square kilometer. Cell densification has been earmarked to generate more capacity gain for 5G networks [50]. 6G is expected to be more heterogeneous than 5G, thus presenting more promising scenes for computation offloading due to the proximity of UEs to SBSs, higher capacity, and lower latency. However, the increase in SBSs raises high energy costs, and the closeness of SBSs to one another can generate severe co-channel interference [2, 51]. Co-channel interference can immensely complicate the computation offloading decision, which is determined by the wireless transmission condition. Hence, the interference mitigation techniques, such as transmission power control, frequency subcarrier allocations [52], adaptive beamforming, interference cancelation, interference randomization [53], and coexisting cloud and edge AI [48] should be considered in the future wireless networks design to improve the rate of offloading tasks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

4.1.3 Bandwidth and Power Allocation

Bandwidth and power allocation has been a pivotal technique of improving network efficiency under guaranteed QoS to users. In orthogonal frequency division multiplexing (OFDM), bandwidth is allocated by converting a wideband spectrum into several narrowband orthogonal subcarrier channels to serve multiple users at a time. That means the subcarrier channels can be shared among several users using the same network concurrently; hence, users can efficiently offload the tasks via the subcarrier channels. Moreover, in multicarrier systems, the total transmitted power (the power required on each subcarrier) is minimized to control co-channel interference while offloading the tasks for computing and caching services. Such technique dramatically reduces interference since a subcarrier can be occupied by at most one user [54]. Thus, the demands for higher throughput, lower latency, and higher reliability will be met. On top of that, 6G holds strong potential to emerge with superior wireless channels, such as universal filtered multicarrier and filtered-OFDM, which could further accelerate the computation offloading efficiency.

4.1.4 Energy Efficiency

The increasing demands for higher communication capacity and fast-growing energy costs make the energy-efficient wireless communication network design an emerging trend. In conventional networks, the radio access part is viewed as the prime energy consumer, accounting for greater than 70% of the total energy consumption [21]. In the design and analysis of green networks, EE metrics are intrinsic since they help assess and compare the consumed energy of various designs and provide long-term research goals [55]. Designing an energy-efficient wireless network is highly desirable for efficient tasks offloading. Hence, EE has to be improved in the future wireless networks design. According to [48], the network EE will enhance to 10-100 $\times$ that of 4G and 10-100 $\times$ that of 5G, in 5G and 6G networks, respectively. Therefore, as a KPI for evaluating 6G wireless networks, EE has to be thoroughly considered.

4.1.5 Achievable Transmission Rate

Several studies explored the maximum limits of the achievable transmission rates of wireless communication networks over fading channels [56]. The study of time-varying fading channels, where fading gains (channel states) are often modeled as stochastic processes, such as independent and identically distributed (i.i.d.) process and Markov process, yield tremendous achievements [57]. For modeling wireless fading channels, finite-state Markov channel (FSMC) is mostly considered. The FSMC model, which relies on partitioning the received signal-to-noise ratio (SNR) into a finite number of states, has drawn much attention due to its great balance between complexity and accuracy [58].

The efforts in [59–61] modeled the wireless channels as FSMC; the goal is to effect higher efficiency than that of traditional assumption of static channels. In such wireless scenario, Shannon's theorem can be used to evaluate the achievable data rate. In other words, in a cellular cell with an available bandwidth allocated to each UE and a given BS transmission power, the ultimate wireless transmission rate of a mobile UE is defined by the Shannon's capacity theorem [62]. Hence, based on the achievable SE ℓ_u^k expressed in (1), the achievable transmission rate of the UE u can be expressed as;

$$R_u^k = \psi_u^k B \ell_u^k, \tag{2}$$

where B represents the available communication bandwidth and ψ_u^k denotes the fraction indicator ($0 \leq \psi_u^k \leq 1$) of the available bandwidth allocated to the UE u.

The time (in seconds) taken to offload a computational task n from a mobile UE u through a wireless channel, i.e., transmission time/delay can be obtained by dividing the task input-data size (denoted by L_n) by the achievable transmission rate expressed in (2). Hence, transmission time does not depend on the channel length, rather, it relies on the input-data size/quantity of the task and the data transmission rate, and it can be given by;

$$T_{u,n}^k = \frac{L_n}{R_u^k}, \tag{3}$$

where L_n (in terms of bit) represents the input quantity of the task n; see [49], [63].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

As per [63], the energy (in joule) consumed while transmitting a task n from a mobile UE through a wireless channel can be obtained simply by multiplying the transmission power of the UE u (p_u^k) and the transmission time expressed in (3). This energy can be expressed as;

$$E_{u,n}^k = \psi_u^k p_u^k \frac{L_n}{R_u^k}. \tag{4}$$

In wireless networks, the uplink/downlink transmission of UE is generally managed by a wireless BS, which might be a Wi-Fi AP, a Femtocell network AP or even a macro-cell BS [64]. RAN is widely seen as an intrinsic part of the wireless network infrastructure facilitating wireless connection between the UE or any wireless controlled device and the cellular core network. In an MEC system, the UE is wirelessly connected to RAN; likewise, RAN is connected to the core network through the guided channels like IP/ethernet. In particular, RAN provides connection between BSs and backhaul networks through the ethernet interface, supporting high-speed data transmission [65].

In mobile cloud computing (MCC), the communication paths between the UEs and cloud servers are normally abstracted as bit pipes characterized by constant or varying rates with given distributions. The point behind such models has to do with tractability and may be connected with an MCC system design, focusing on handling the latency issues in the core network and managing the large-scale cloud. Thus, these bit-pipe channel models do not pay attention to the wireless communication network latency. Conversely, in a small-scale edge cloud, such as an MEC system, the focus is to dramatically minimize the communication network latency through an advanced air interface design. Moreover, the bit-pipe models omit some important features of wireless propagation and are too simple to support implementing modern communication techniques [7].

4.2 Computation Model

4.2.1 An Overview

Computation offloading associates with communications through two major issues, i.e., latency and energy consumption. Nonetheless, it plays a pivotal role in computation aspects [66]. In literature, several offloading destinations were considered for executing the computational tasks. The main destinations include: i) cloud-only, in this class, the computation-intensive tasks and latency-tolerant applications are pushed to the remote cloud server for execution, ii) MEC-only, herein, the latency-sensitive applications, e.g., VR, Tactile Internet, face recognition, and other mission-critical IoT applications, are migrated onto an MEC server for speedy execution, and iii) local, where the tasks are rather executed locally at UEs than at an MEC/cloud server due to higher energy consumption and latency [67]. In [68], three key steps were identified for offloading computation-intensive tasks from UEs; these steps are: i) tasks partitioning, ii) offloading preparation, and iii) offloading decision, as shown in the sequel.

Tasks Partitioning: The tasks here are grouped into offloadable and non-offloadable components (tasks). The former involves the tasks that should be migrated for remote execution; whereas the latter refers to the tasks that should be retained at UEs. Identifying computation-intensive tasks to be offloaded requires performing source code analysis and performance prediction by an application programmer [68]. Determining part of the tasks that should be offloaded for execution is important in partial offloading; thus, there is a need for partitioning tasks into modules. The partitioning approach can either be static (where the tasks are partitioned into a fixed number of partitions during application development) or dynamic (where the tasks are partitioned at runtime based on the availability of bandwidth and quality of network connection). The static is quite easier to implement. However, the dynamic is important in cases where the computation-intensive applications are required to adapt to the network and mobile environment changes. For instance, some frameworks, such as MoSeC and Self Cloning, employ dynamic partitioning; whereas others like Aura, Avatar, and MALMOS use both static and dynamic [69].

Offloading Preparation: This covers all essential steps, such as remote server selection, migration, code installation, and tasks or data migration for remote computing, required for offloadable tasks to enable their use in native UE applications [68].

Offloading Decision: This usually precedes remote execution of offloadable tasks. The execution context

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

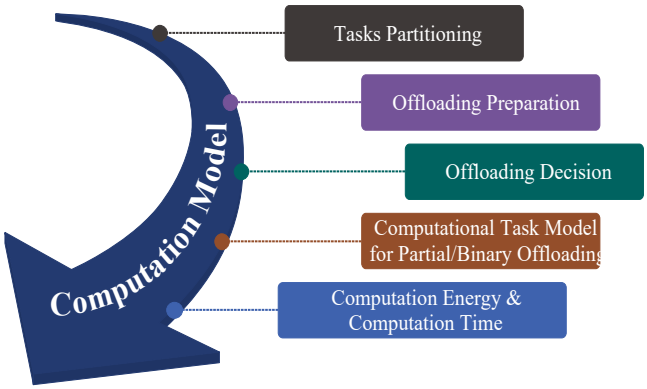

Figure 5. *A rodmap of the computation model.*

determines whether an installed remote task is used in the UE applications or not [68]. Offloading decision is an extremely complex process affected by various factors, including subscriber's preferences, application (nature), UE capability (e.g., high or low performance UE), application model, connection (e.g., network bandwidth, delay, and costs), and cloud/MEC service [70, 71]. Generally, offloading decision is carried out in sequence. This can be explained in an MEC environment, in such scenario, UE decides whether to offload a computational task via wireless channels for remote execution or execute them locally. If the task is offloaded to the network edge/MEC server, the server determines whether it can meet the request or the computation should be further offloaded to the remote cloud for execution. Offloading decision can either be full or partial. The gains derived from full offloading decision are to: reduce the consumed energy at UEs while satisfying the constraints of delay, ii) lower the computing delay, and iii) determine a decent tradeoff between the computing delay and energy consumption. In partial offloading, the benefits are to: lower the consumed energy at UEs while satisfying the constraints of computing delay and ii) find appropriate tradeoff between the computing delay and energy consumption [71].

Before walking through the computation model, we find it helpful to briefly look into partial offloading and binary offloading at this point. The former enables partitioning of computational tasks into different parts at UEs for local computing and offloading simultaneously. Whereas the latter makes it impossible to partition the computational tasks; rather, the tasks are executed as a whole either at the UEs or at the network edges [72]. To simplify the description of the remote CPU computation model, we restrict the scope of our discussion to the computational task model for binary offloading, which is subsequently followed by computation energy and computation time as following.

4.2.2 Computational Task Model for Binary Offloading

Here, the task model can be described in terms of two or three different fields. In a two-field task model, a computational task n can be characterized by $T(L_n, D_n)$. In this case, $L_n > 0$ (in bits) and $D_n > 0$, respectively, depict the task input size and the total number of the required CPU cycle to execute the task [64]. The size of the computation output is usually assumed negligible; thus, it was ignored in [64], [73–75]. There are various factors on which the required CPU cycle number depends for executing the tasks. These factors include specific applications, task input size, and physical components, such as memory and CPU, in the computing device [73]. However, the number of the required CPU cycle is not considered in some two-field notation models; rather, a task completion deadline might be a priority especially when computation delay becomes a concern. Such model was adopted in [73–75], where $T(L_n, \tau_n))$ characterizes the computational task n. This implies that each task n with an input-data size L_n has to be accomplished within its completion deadline, denoted by $\tau_n > 0$ duration. For the latency-sensitive applications, it is assumed that the completion time cannot exceed the coherence time of the channel. Thus, the channel power gain does not change within the block of interest.

The computational task requested by the UE applications can be modeled as i.i.d. Bernoulli process. To be specific, for each time slot, the probability of requesting a task can be denoted by ρ, and that of not requesting it can be given by 1-ρ. Such request is often followed by an execution decision [74]. For instance, when a mobile VR device requests a computational tasks, an MEC server determines whether the requested tasks should be computed on it or not. If the tasks should not be executed at the server, the tasks or their corresponding parts (chunks) have to be forwarded to the VR device for local execution [76].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

Alternatively, the three-field task model has been adopted in several efforts, including [7], [63, 77]. In such model, a three-field notation $T(L_n, X_n, \tau_n)$ can be applied to represent a computational task n. This notation carries the information of the task input size $L_n > 0$ (in bits), the computation workload/intensity $X_n > 0$ (in CPU cycles per bit), and the task completion deadline $\tau_n > 0$ (in seconds) [63, 76, 77]. The three fundamental parameters of a task are determined by the nature of the task itself [7, 63, 77, 78].

The CPU workload of a computational task directly determines the consumed energy of computing. The workload is determined by the total amount of the required CPU cycles. In short, the relationship between the number of the required CPU cycle (D_n) for computing a task n and the task input size (L_n) with the computation intensity (X_n) can be expressed as: D_n= $L_n X_n$. Hence, the number of the required CPU cycles for executing computational tasks differs in various applications, and it may be determined by offline measurement [74].

4.2.3 Computation Energy

The power consumption in a CPU depends on a number of factors, such as short circuit power and dynamic power. However, the main energy consumer is dynamic power, which can be controlled by adjusting the CPU-clock frequency of the chip voltage based on the dynamic voltage and frequency scaling (DVFS) mechanisms [75]. The DVFS varies the CPU supply voltage and clock frequency of UEs according to computation load to meet the performance requirements. With the DVFS mechanisms, the CPU-clock frequency of UEs will be regulated to lower the energy consumption in an adaptive way. Therefore, by incorporating DVFS techniques into computation offloading, the strategy design becomes more flexible [79].

For the computational task n, the total energy consumption constitutes the energy consumed while offloading the task from the UE to the network edge (e.g., SBS k) and the energy consumed while computing the task at the SBS. This is given by$E^k_{u,n}$ (as expressed in Section IV-1) and $E^k_{u,n}$, where E^k_n= $\mu_0 f^2_k D_n$, μ_0 is a constant related to the CPU of a computing server in the SBS k, and f_k denotes the CPU cycle frequency of a computing server in the SBS k. Here, the computation result is assumed negligible [63]. To that effect, the total energy (in joule) expended for executing the task n at the network edge can be represented as

$$E_t = E^k_{u,n} + E^k_n. \tag{5}$$

4.2.4 Computation Time

Here, our focus lies in the total time costs for executing a specific task at a network edge, consisting the time consumed while transmitting the task to the network edge and the actual time spent while executing the task at the edge. Thus, a computational task n, denoted by $T(L_n, X_n, \tau_n)$, can be offloaded through an OFDMA channel with maximum achievable transmission rate to the network edge (e.g., SBS k). In this regard, the time expended for offloading the task n to the SBS k can be given by $T^k_{u,n}$, as expressed in Section IV-1. Likewise, the time cost for executing a task at the network edge, as defined in [63], refers to the total amount of the required CPU cycles for computing the task divided by the corresponding CPU-cycle frequency. Hence, the time consumed for computing the task at the SBS k, which depends on both CPU-cycle frequency f_k and computation intensity, can be given by $T^k_n = \frac{L_n X_n}{f_k}$; see [63, 80, 81]. To that effect, the total time cost (in second) for executing the task n at the network edge can be expressed as

$$T_t = T^k_{u,n} + T^k_n. \tag{6}$$

4.3 Caching Model

4.3.1 An Overview

In the previous decades, research studies devoted to CDNs gained momentum by focusing on where to deploy the servers (server placement); which files to cache at each server (content placement); how much storage capacity to allocate to each server (cache dimensioning); and how to route content from caches to end-users (routing policy) [82].

The caching systems may differ in terms of granularity, scale, and technologies. Nonetheless, the common goal shared by all caching systems is to optimally cache data contents for subsequent usage [83]. To provide clear description of the caching model, this overview touches on the two basic approaches to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

caching studies, caching places, and performance of cache networks.

Approaches to Caching Studies: The effort in [84] described two basic approaches to caching studies, i.e., coded and un-coded caching. One of the approaches involves the conventional caching schemes, such as first-in-first-out (FIFO), least frequently used (LFU), least recently used (LRU), etc. Such schemes generally use cache hit ratio as a key parameter for performance evaluation and are often called un-coded caching schemes because there is no coding in them. Besides, each of the schemes is featured with particular insertion and eviction policies. The other approach, i.e., coded caching, involves content placement and content delivery. In the placement phase, caches are populated with file contents usually during low network activity, i.e., off-peak hours. While in the delivery phase, the server serves the requests by executing coded multicasting. The peak/average number of file contents transmissions via the shared link is commonly considered as a key performance metric. Since coding is used for delivering content, this approach minimizes the file transmissions within the networks and attracts extensive research investigations; see [85, 86].

Caching Places: In mobile networks, several places can be considered for deploying edge servers and content caching. The three key places where cache can be deployed to cache file contents in cellular networks constitute UEs, RAN, and the core network [14, 87]. Since deploying cache at the evolved packet core (EPC) is technically more convenient than deploying cache at RAN, EPC is considered as the most commonly deployed caching place. At the network edges, content may be cached in MBS, SBS, or UEs [14]. In MEC, caching content at the network edge brings significant gains, enabling MEC to get real-time information from RAN and utilize it for guaranteeing QoE of UEs. Hence, with real-time information, remote/MEC server optimizes the users' traffic to ensure QoE [88].

Performance of Cache Networks: The main considerations for overall caching performance involve caching policies, deciding what to cache and when to deliver the caches [89]. Hence, user's request patterns, caching policies, and how caches are operated (cooperatively, independently, or in a globally coordinated manner) constitute some of the several factors on which the performance of cache networks depends. The caching policy is vital; upon a user's request for data, it makes decisions whether to cache the data, where to cache it, what timer value to set in case of using time-to-live caches, or which data to evict in case of a full cache [90]. Various caching policies have been offered in literature for managing a single cache, which differ in terms of either eviction or insertion policy. In [91], some existing caching policies, such as FIFO, LFU, LRU, q-LRU, k-LRU, RANDOM, and k-RANDOM were studied.

Being easy to implement, LRU is widely adopted and it provides great performance. In the context of ICN, FIFO and RANDOM are considered as feasible substitute for LRU since their hardware implementation in speedy routers is easier. The k-LRU and q-LRU enhance the LRU performance through advanced insertion policy; see [91]. Estimating the gain behind a content by assessing its present popularity, potential popularity, cache capacity, and locations of existing replicas over the network topology is essential. Instead of applying conventional policies, including FIFO, LFU, and LRU, it is desirable to propose cooperative caching policies for EPC and RAN caching in order to efficiently increase cache hit rate [89].

The concept of caching has been advancing rapidly due to the unprecedented growth of network traffic over wireless networks. With the advent of MEC, both caching [92] and computing are pushed near the UEs, i.e., network edges. This implies dramatic reduction in both content delivery delay and network traffic and also guarantees adequate computing and caching functionalities. Indeed, efficient caching functionality at the edges allows mobile UEs to alleviate the possible burden from backhaul links. Thus, caches can be designed for efficient communication between edge servers and UEs [88], [92]. Moreover, the classification of caching techniques, not only in the context of MEC, can be either reactive/transparent or proactive. In transparent caching, neither UE nor the application service provider (ASP) is aware of a caching MEC server. In proactive caching, data contents are non-transparently cached before being requested since it can result in high network utilization in future [93].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

4.3.2 Cache Performance Metrics

In conventional caching scheme, cache-hit rate/cache-hit ratio is considered as a common performance metric, representing the ratio of the requests satisfied by a caching system and the aggregate incoming requests. Generally, high cache-hit ratio implies a high-performance caching system since it brings about dramatic reduction in redundant data transit. There are four factors on which cache-hit ratio depends, i.e., cache size/capacity and cache algorithm; these two can be figured and controlled to a certain extent. The other two factors are content population and content popularity distribution; these cannot be controlled. They are externally generated by subscribers and corresponding applications that interact with the caching system. Despite being widely viewed as a common performance metric of caching systems, cache-hit ratio is incapable of providing insights into the performance of network of caches. Thus, being a conclusive result of request filtration that all caching systems in a network achieve, server hit ratio is considered as a more appropriate metric. Another KPI is footprint distance. Here, the shorter the footprint distance, the nearer the data content is to the requesters, hence implying shorter response time for the requesters. In short, for green communication networks, caching time is an essential index. Hence, in designing a cache algorithm, footprint distance, server hit ratio, and caching time have to be considered [83].

4.3.3 Cache Capacity

The cache capacity or size of cached information (in bytes) is widely considered as the typical measurement of caching capability [13, 46]. With increase in the capacity of cache memory, the cache-hit ratio can be improved. Compared with content population, the cache capacity is relatively small. However, multi-magnitude increase in the cache capacity may lead to a few percentage points of cache-hit ratio improvement [84]. At the network edge, after computing the tasks, the outputs might be considered reusable. In such case, a BS with a given caching capacity, say C bytes, can be required to cache the outputs [78]. Therefore, to cache content at the network edge, each BS decides whether the content offloaded from UEs should be cached in its cache memory before or after computation based on each content's popularity distribution. This is achieved by considering two binary parameters to control the caching strategy, e.g., $\gamma_u^{k,1}$ and $\gamma_u^{k,2}$. If it is decided that BS will cache the original content $\gamma_u^{k,2}$ is set to 1; otherwise, it is set to 0. On the other hand, if it is decided the BS will cache the computed content $\gamma_u^{k,2}$ is set to 1; otherwise, it is set to 0 [94]. The cached content may be represented by a set of finite numbers, where each content can be a short video clip or a portion of movie with a given size (in bits) [95].

4.3.4 Cache Capacity Constraint

Since the capacity of cache memory of a BS (e.g., MEC server) is finite, the total size of cached content cannot exceed it [94]. That is to say the caching capacity constraint, expressed in (7), has to be satisfied. Specifically, for a given cached content, the caching capacity constraint can be expressed as

$$\sum \gamma_u^{k,2} L_r \leq C, \tag{7}$$

where C (in bytes) denotes a given caching capacity and the cache decision variable and computation result can be denoted by $\gamma_u^{k,2} \in \{0, 1\}$ and L_r, respectively [78, 96].

In the conventional (un-coded) caching scheme, the gain is derived from making content available locally. Actually, a UE may request some content cached in its cache; in this way, the local memory of the UE serves this request. This implies the local caching gain, which is essential if the local cache memory is big enough to cache ports of the popular content locally. In the coded scheme, as per [76, 77], the global gain is derived from joint optimization of the placement phase and delivery phase, ensuring that various requirements are met in the delivery phase with the single coded multicast transmission. Because content placement is carried out without knowing the actual requirements, realizing the global gain requires careful design of the placement phase such that multicasting opportunities can be created at the same time for all possible requests in the delivery phase. In a word, if the aggregate global cache capacity exceeds the total content size, then the global caching gain becomes relevant [29, 85, 86].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

4.3.5 Caching Reward

In mobile wireless networks, the reduction of the network backhaul delay or the backhaul bandwidth alleviation is considered as a caching reward. Hence, the reward of caching content requested by a mobile UE can be expressed by $\gamma_u^{k,2}Rc$. In this regard, c and R, respectively, represent the content request rate requested by a mobile UE u and the average data rate of a single UE in the system [96].

4.4 Control Model

4.4.1 Distributed Control Model

Designing an accurate control model that controls, coordinates, integrates, and optimizes communications, computing, and caching resources can be highly complex. Recently, [19] adopted a distributed control model based on a distributed optimization by which the communication, computing, and caching models can be coordinated and integrated at the network edges. The distributed control model enables the MEC servers hosting computing and caching resources to be deployed at the same domain and collaborate to share resources. In this way, the information exchange between the MEC servers and centralized cloud server can be reduced, thereby minimizing the backhaul bandwidth. Hence, with the distributed control model, the cache hits can be improved. As discussed in Section IV-3, the reduction of backhaul delay is termed caching reward. Here, the amount of backhaul bandwidth saved by the distributed control is adopted as a caching reward.

Moreover, the distributed control enables the exchange of a small amount of information among the collaborating MEC servers, thus bringing significant gain in terms of maintaining the resource allocation in the vicinity of accessible computing and caching resources. However, among the collaborating MEC servers, there is no provision for a centralized controller that controls the entire servers. In such a case, the distributed control can be modelled as dynamic feedback control model; see [97]. So then, the resource allocation table update at each server can serve as a feedback with a given state at iteration t, which is used for determining the new state at the next iteration $t+1$; see [19].

Therefore, instead of collecting all problem parameters and performing a central calculation, several agents, obtaining certain problem parameters by sharing information with finite set of neighbors, compute distributed algorithms. The agents may represent BSs, edge/MEC servers, UEs, or buses depending on the specifics of the distributed algorithm and the application of interest. Compared with centralized approaches, distributed control algorithm brings several gains. For instance, the computing agents can only exchange small amount of information with a subset of the other agents. This means the expense of the challenging communication infrastructure can be lowered and also cybersecurity can be improved. Due to its ability to do parallel computations, distributed control algorithm can be computationally superior to centralized control algorithm when it comes to the maximum problem size that can be handled and solution speed. In addition, distributed algorithm is robust in terms of failure of individual agents. Finally, distributed algorithm has the potential to respect data privacy, cost functions, measurements, and constraints, which becomes more significant in a distributed generation scenario [98].

4.4.2 Hierarchical Control Model

In hierarchical modelling, models may represent different parts of a studied system or its various properties that are logically ordered to form a hierarchy or a sequence. In modelled systems description, the lower hierarchical levels usually correspond to higher levels of detail. Besides, there is nearly similar level of detail in each element of a sequence, and the outputs of a present model imply the input data of a succeeding/next model [99]. Molzahn et al. [98] described hierarchical control scheme as algorithms where computations are performed by agents that exchange information with other agents at a higher level in a hierarchical structure, eventually leading to a centralized control. The promising gains of hierarchical control triggers extensive research efforts.

In [34], hierarchical control scheme was proposed to model the interactions between SBSs and UEs by integrating bandwidth allocation for communication, computation offloading, and cache splitting. The control scheme harnesses the synergistic combination of caching, computing, and communication capabilities in the SBS, hence characterizing competitive and col-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

laborative interactions among them. Specifically, a two-tier hierarchical game model was applied based on a unified and integrated approach to model the interplay between the SBSs and UEs. The control decisions are made by the game players, i.e., SBSs and UEs, in line with the step-by-step timed learning approach. In the first-tier, only SBSs act as the game players. The communication bandwidth is shared among these SBSs as per a dynamic bargaining model. In the second-tier, each SBS and its respective UE act as the game players, and the interactions among them are modelled as Stackelberg game model. The SBS, serving as a leader, splits its caching capacity and decides the cost of communication and computing services. The leader's decision is monitored by the UEs (followers), which select their appropriate strategy. In [19], the hierarchical control enables offloading decision-making at UEs and enables each MEC server, as a controller, to decide for the offloaded tasks. In a word, control entails decision-making with respect to what, how much, where, and when to allocate available resources to a given task.

However, limited resources and heterogeneous users' demands make resource allocation a complex problem. Generally, the types and amount of the required resources are determined by users in their requests. Network service providers, in response, allocate the resources requested while ensuring that the granted resources are sufficient to meet the constraints defined by the users. As a result, any control strategy for resource allocation systems i) has to be focused on requested resources, ii) must satisfy the users' ever-changing demands, iii) has to be optimal by ensuring optimal utilization of resources, and iv) prioritize task for superior performance [100]. These requirements make resource allocation system control very challenging and thus limited. Hence, coming up with tractable control techniques that guarantee optimal distribution/allocation of a system's resources is not an easy task. Nearly all the optimal or near-optimal resource allocation system control techniques/algorithms are not tractable. On this account, researchers make a trade-off between tractability and optimality. Recently, Lima et al. [4] achieved near optimality of control system and stability by employing the constrained reinforcement learning (RL) techniques. This achievement portrays strong potential of intelligent control in overcoming the shortcomings of conventional resource allocation control techniques.

4.4.3 Intelligent Control Model

This involves learning, decision-making, and optimization based on the frontiers of AI, e.g., DL, ML, and DRL. Intelligent control relies on utilizing existing knowledge or experience to enable various agents to intelligently learn, optimize, and take appropriate actions (e.g., resource allocation control, network association, and resource management) with dual functions for supporting diversified network services. Such functions can be realized with AI techniques applied in 6G networks. Thus, network agents, such as BSs, MEC/edge servers, and UEs, can be equipped with learning models (intelligent brain) such that they automatically learn to make resource allocation decisions [10]. Generally, edge devices or edge servers host both edge resources and training at the network edges. Such servers are not powerful as computing clusters or centralized cloud servers. Xu et al. [12] raised four major problems that need consideration for edge training, i.e., i) how to train (the training architecture), ii) how to speed-up the training (acceleration), iii) how to optimize the training approach (optimization), and iv) how to assess the vulnerability of the model outputs (uncertainty estimates).

Today, the intelligent control brings striking gains in terms of coordinating, controlling, and optimizing mobile communication, computing, and caching resources. Considering the heterogeneous nature of both wireless networks and UEs with challenging QoS requirements of the UEs applications, the conventional resource allocation optimization and control algorithms cannot be sustainable for performance requirements of 6G and beyond mobile networks. Hence, the promising answer lies in the AI-based control (which can be distributed algorithms/model), which has been attracting extensive studies.

To realize the i4C framework, these models can be jointly optimized by considering the possible constraints, decision variables, and optimization objectives. Depending on the specifics of the optimization and the application task of interest, i) the decision variables comprise the computation offloading, data caching, execution, and resource allocation decision variables; ii) the optimization goals could be network performance, costs of deployment and opera-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

tion (communication/computation costs, energy consumption, etc.), efficiency, network/system reliability, and privacy [42]; and iii) the constraints may include computation time, local computing capabilities of UEs, caching and computing capacities [19], wireless channel states, dynamic trust values, cache status [60], computation delay [73], and so on. Thus, various optimization techniques can be considered to realize the i4C model.

Note that, some traditional optimization techniques may not yield desirable results due to the numerous configurable parameters, wireless channels conditions, multiple decision-making variables, and heterogeneous users' demands; besides, the complexity of mobile networks is still growing. However, among the conventional optimization approaches, the Lyapunov optimization method excels as the best candidate for the long-term stability of dynamic systems. What follows in the next section is devoted to various optimization approaches for achieving the i4C in 6G and beyond networks.

V. INTEGRATION OF 4C

This section presents a great deal of research efforts aimed at combining 4C. We observe that several cutting-edge research efforts on the i4C focused on applying the AI techniques to integrate/optimize 4C at the network edges due to the growing complexity of mobile wireless networks. This is in contrast to earlier efforts that tended to integrate/optimize the resources based on the conventional resource allocation optimization approach. This section considers both conventional and recent integration approaches. Specifically, the section reviews recent works on the i4C based on the conventional optimization approach in Section V-1, discusses intelligence for resources integration in Section V-2, reviews recent trends in AI-based integration approach in Section V-3, brings a snapshot of the various approaches devoted to the integration of 4C in Section V-4, and discusses the integration of sensing and communication in Section V-5.

5.1 Conventional Approach

Now that it becomes a common fact that the network resources/functionalities are on the verge of attaining their maximum performance, integrating them becomes unavoidable. Such integration will guarantee true pervasiveness. In future, rather than subscribing to individual services separately, subscribers will most likely subscribe to a service provided by integrating several services in different domains, which may include communications, caching, computing [101], and control. The confluence of these resources will induce a remarkable transformation in the design philosophy of future networks [14]. Hence, combining 4C in the overall 5G, 6G, and beyond mobile networks design becomes necessary, especially for surmounting the challenges of emerging technologies.

To this end, the efforts in [19] and [34] focused on utilizing the network edges to achieve maximum network utility through the convergence of 4C. In particular, Ndikumana et al. [19] proposed an integrated framework of 4C for managing big data in MEC. The framework enables big data computing and caching operations at the MEC servers, thus reducing end-to-end latency. These servers occupy the same cluster and actively collaborate to share the 4C resources. The technical rationale behind this is to: i) improve the backhaul network traffic, ii) maximize utilization of resources, and iii) minimize latency in the integrated 4C. The authors collaboratively optimized 4C to maximize bandwidth while minimizing latency under the constraints of the computing deadline, the local computational power of UEs, and MEC resources. Because of decision-making variables at multiple locations, the optimization problem, which is non-convex, was solved using a modified version of the block successive upper bound minimization (BSUM) approach.

In contrast, Kim [34] applied a hierarchical game-theoretic control algorithm for integrating resources. To be specific, the author devised a 5G network SBS, where a holistic control scheme characterizes the competitive and collaborative interactions among the communication, computing, and caching resources. This was realized by adopting game theory, which has to do with tactical interactions among several intelligent logical decision makers that systematically follow their objectives while maximizing the anticipated value of their payoffs. Due to its inherent ability to define the interactions among intelligent agents with conflicting objectives, game theory is applied to handle diverse competitive issues of network resources in wireless communications. Thus, the proposed approach considers: i) offloading decision at

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

each UE, ii) radio splitting decision for data and content caching capacities, and iii) bandwidth allocation decision for each individual SBS. These decision issues require leveraging design principles, including self-interactivity, feasibility, and integral combination of various control algorithms, which depend on each other, to deal with conflicting performance benchmark under highly diverse 5G network circumstances.

In a design of 5G network SBS, where data caching, computation offloading processing, and mobile communication technologies are jointly utilized, a new control paradigm has to be employed to leverage the synergistic benefits of the integrated communication, computing, and caching operations in the SBSs. By doing so, different communication, computing, and caching characteristics can be captured to realize a promising solution under diverse 5G network circumstances. Therefore, as a control theory of several goal-oriented agents, game theory offers numerous promising solutions that optimize the overall performance of 5G networks; see [34].

In [102], Huo et al. relied on the principle of programmable control and caching, stemmed from SDN and ICN, respectively. Based on these premises, they proposed an integrated framework that systematically combines in-network caching, computation, and networking resources. These resources are centrally controlled and managed by utilizing SDN controller, thus monitoring both UEs and resources in the data plane. The framework enables dynamic orchestration of networking, caching, and computing resources to match the needs of future green networks. Conversely, [35] employed SDN to introduce an integrated framework of 4C. By fully considering the ability of the programmable control in SDN, Chen et al. proposed a framework that systematically converges networking, computation, and caching resources and allows the control functionality to dynamically orchestrate them. Unlike SDN, which programmably controls the switching devices' forwarding function, this framework controls the data plane's three-dimensional resources. It is also considered as service-oriented that supports general in-network services, thus different from content-oriented ICN with fixed in-network services. The study in [38] described such framework in Fig. 6, where the data plane is composed of caching, computing, and forwarding devices; the management/-control modules for computing and caching resources

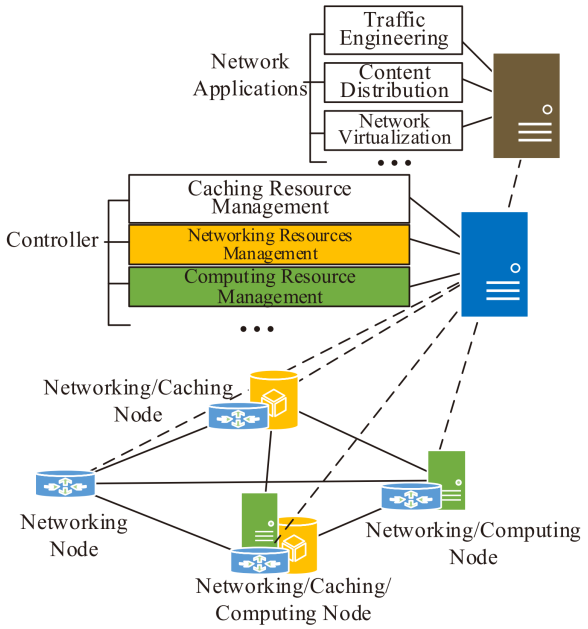

Figure 6. *An integrated 4C framework [38].*

are appropriately deployed at the control plane. The management plane bears the responsibility of monitoring and configuring the control functionality remotely by leveraging the SDN controller. The framework realizes an effective resource allocation and network orchestration by dynamically guiding various computing and content services to the corresponding service requesting UEs.

5.2 Intelligence for Resources Integration

With the recent emergence of edge intelligence, intelligent control will inevitably interact with computing, communication, and caching functionalities at the network edge to promote the overall network performance and maximize end-to-end users' experience. Currently, there are two key factors in existing literature that motivate integrating/optimizing the edge resources with intelligent control, i.e., 1) complexity of mobile wireless networks and 2) exponential growth of data, as explored in the sequel.

5.2.1 Complexity of Mobile Wireless Networks

Mobile wireless networks are becoming more complex nowadays. Such complex networks involve networking, computing, wireless communications, caching, and so on [44]. Moreover, the incoming

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

6G networks will most likely be highly complex and dynamic. Conventional optimization algorithms with weighty mathematical models, such as gradient methods and Lagrangian duality, may not be desirable candidates for 6G. Hence, in AI-powered 6G networks, the parameters and frameworks will be optimized by utilizing the AI techniques rather than conventional tedious computation. Such techniques have good prospects of training and self-learning models for realizing network optimization in 6G networks, enabling network operators and providers to optimize the resources and parameters for enhanced QoS [10].

Wang et al. [59] raised some points spurring the convergence of intelligence with the communication, computing, and caching functionalities. Specifically, the effort pointed out that several existing efforts handled the resource allocation issues. Such efforts, in their assuming settings (mostly built upon game theory, convex optimization, etc.), realized quite impressive results. However, with specific use cases in MEC, such optimization approaches can be limited by: i) *Uncertain Inputs*: here the assumption is that some relevant information factors are considered as inputs, unfortunately obtaining some of them remains a herculean task owing to randomness in wireless channels and privacy policies; ii) *Dynamic Conditions*: the confluence of communication and computing resources dynamics are still under study; and iii) *Temporal Isolation*: apart from Lyapunov optimization, most of those optimization problems overlook the lasting effect of current decisions on the allocation of resources, i.e., majority of traditional optimization algorithms are near-optimal or optimal only for a snapshot of the system in an extremely time-varying dynamic MEC framework. In short, diversified network devices requirements, wireless channels variations, different QoS needs, and much more complicate conventional optimization of 4C in 5G and 6G networks [8].

Besides, edge computing requires powerful optimization tools to address resource allocation challenges at various layers, including bandwidth, radio frequency, access jurisdiction, cache capacity, and CPU cycle frequency. AI solutions will surely handle such tasks [42]. Therefore, in 6G and beyond networks, a finest solution to the resource allocation optimization is attributed to AI (e.g., ML, DL, and FL techniques). Deng et al. [42] highlighted how to utilize the AI techniques at the edge to obtain more optimum solutions.

5.2.2 Exponential Data Growth

Today, we witness the unprecedented growth of data, created by different prevalent devices from mobile phones to industrial robots [41]. In fact, billions of connected IoT and mobile devices generate massive volumes of data in zillions of bytes at the network edges. Motivated by this trend, the AI techniques have to be pushed to the network edges in order to completely unleash the capability of the edge big data [103]. Hence, by leveraging an MEC platform and utilizing voluminous data distributed over myriad connected devices, the limitations of computing capability and finite data can be overcome at each device. The two critical and combined aspects in such systems are i) communicating between connected devices and edge servers and ii) learning from distributed data [104]. Learning involves harnessing or altering existing knowledge and experience with the goal of improving a device or service center. Learning will bring numerous gains to 6G networks (including resource management, optimum network slicing, and so on), as per the applications requirements [10].

At the network edges, AI can exploit its incredible learning and reasoning capability to obtain significant information from data and perform decision-making, thus realizing intelligent management and maintenance of a network involving communication, computing, and caching resources [44]. In a word, big data analytics and real-time processing require the distributed integration/optimization of the network edge resources and intelligent control at close proximity of the data sources.

5.3 AI-based Approach

The thrust of recent studies in [59, 60] and [105–108] focused on overcoming several identified issues of the i4C by harnessing the AI frontiers at the network edges. Wang et al [59] examined resource allocation optimization constraints in time-varying MEC systems and accordingly incorporated the DRL techniques (explicitly, Deep Q-Learning) and distributed DRL approaches into mobile edge systems. By coupling the DRL techniques and FL framework with MEC, the authors designed a framework to optimize communication, computing, and caching resources with intel-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

ligence at the edges. The 4C framework, termed In-Edge AI, intelligently harnesses the collaboration between the edge nodes and UEs to ensure better training and inference of the models through the exchange of learning parameters. Thus, performing dynamic network-level optimization and application-level enhancements meanwhile minimizing unnecessary communication burden. The gains are i) with DRL at the edges, information collecting, cognitive computing, and requests handling can be enabled, leading to an effective management of joint resources and ii) with FL framework, the DRL agents can be trained in a distributed way while: dramatically minimizing the quantity of data that should be uploaded over the wireless uplink channel, protecting the individual data security, responding cognitively to wireless networks' conditions and mobile communications environment, and fitting well with multiple mobile UEs in an effective mobile network.

He et al. [60] focused on studying the dynamic nature of communication, computing, and caching resources and how to optimally allocate them using automatic decision-making intelligent control. The central idea is enabling efficient resource sharing for mobile social networks under the integrated resources scheme. In particular, the authors explored the sharp increase in trust-based social networks with the recent growth of D2D communications, MEC, and caching and proposed an integrated resources scheme. The integrated scheme faces resource allocation complications for UEs, especially in time-varying network resources conditions. Thus, when the dynamic trust values, wireless channel conditions, computational capabilities, and the cache status are jointly considered, the integrated network becomes more complicated, hence solving the problem by conventional optimization approach will be tedious. In such case, applying an intelligent control, based on the DRL approach, to automatically make resource allocation decisions by considering the networks conditions is a desirable solution. To that end, a Q-learning-based resource allocation strategy was applied to solve the optimized 4C problem without any explicit assumptions or simplification. In [105], Li et al. considered decent gains of RL in time-varying dynamic systems (such as multi-user MEC) and accordingly developed an RL-based optimization framework for handling the resource allocation in MEC. Specifically, the computation offloading decisions and resource allocation were jointly formulated as an optimization problem, which was solved by utilizing Q-learning and Deep Q Network (DQN).

On the other hand, the need to process complex tasks, e.g., airborne/aerial imagery, precision target identification, and adaptive cruise, in the air garners much attention today. Unfortunately, multiple factors complicate the real-time interactions between an MEC system and a UAV. For example, UAV has a short battery lifespan, which may drain fast while interacting with MEC; it may also suffer from insufficient computing and storage resources due to its miniaturized body. Considering these factors, Hu et al. [106] proposed a framework to collaboratively optimize the communication, computing, and caching resources and the UAV swamp with AI-based decision-making scheme. To efficiently process complex tasks in the air by enabling flexible interaction between MEC and UVA, the framework considers UAV swamp deployments both in real-time based on real-time perception and in advance based on historical data mining. The swamp forms a dynamic resource pool based on UAV collaborations in multi-task-offloading scenarios. At the same time, their resources can be dynamically and flexibly allocated to one another so as to balance the resources utilization.

The efforts in [107] and [108] focused on addressing wide-ranging issues of vehicular networks by optimizing 4C functionalities. In [107], He et al. optimized 4C to fulfill the requirements of emerging vehicular networks. Specifically, they introduced an integrated system to dynamically orchestrate networking, computing, and caching functionalities. This was realized based on the principle of programmable control and the concept of information centricity, derived from SDN and ICN, respectively. Then, resource allocation strategy was constructed as an optimization problem while considering the respective gains of computing, caching, and networking functionalities. Due to the high complexity of the system, the authors observed striking features of RL and offered a novel DRL technique for the optimization problem. In [108], the hard service deadline constraints and the vehicle's mobility were considered to design the resource allocation policy. To be more specific, the authors jointly optimized communication, computing, and caching design problem to maximize the cost efficiency and real-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

Table 2. *Comparison of existing works on the i4C resources*

Themes	Networks	Key Contributions	Objectives	Evaluation Method	Related Papers
i4C	Wireless Het-Nets	Proposing an integrated mechanism of 4C for processing big data in MEC.	To increase bandwidth saving and lower latency.	Simulation.	[19]
	5G mobile network	Introducing an integrated control scheme to harness synergies among the mobile communication, computing, and caching capabilities.	To maximize the 5G network performance.	Numerical and Simulation.	[34]
	Software defined networks	Proposing an integrated software defined networking, caching, and computing (SD-NCC) framework.	To promote the system performance and satisfy the needs of diverse applications.	Simulation.	[35]
	Green wireless networks	Integrated framework for dynamic orchestration of network resources.	To satisfy the QoS requirements of various applications.	Simulation.	[102]
	Mobile UAV networks	Introducing an AI-based decision-making architecture for collaborative optimization of 4C and UAV team.	To maximize the network performance by utilizing the available resources.	Experiments.	[106]
	Mobile networks	Integrating the DRL techniques and FL framework with mobile edge systems to optimize communication, computing, and caching resources with intelligence at the edge systems.	To promote content delivery and increase mobile QoS.	Experiment and Simulation.	[59]
	Mobile wireless networks	Introducing a trust-based social networks framework with D2D communication, in-network caching, and MEC.	To maximize the efficiency and security of mobile social networks.	Simulation.	[60]
	Wireless Het-Nets	Integrated system for multi-user computation offloading and resource allocation.	To minimize the total delay and energy consumption.	Simulation.	[105]
	Vehicular networks	Integrated framework for dynamic orchestration of networking, computing, and caching resources.	To increase the performance of future generation vehicular networks.	Simulation.	[107]
	Vehicular networks	Developing a joint optimal resource allocation framework for vehicular networks.	To realize operational excellence and increase the cost efficiency in vehicular networks.	Numerical.	[108]

ize operational excellence of vehicular networks. The formulated problem was solved by developing deep Q-learning based algorithm with multi-timescale framework. Furthermore, the mobility-aware reward estimation for the large timescale model was proposed to reduce the complexity caused by the large action space. Table 2 compares the contributions of existing efforts on the i4C.

In summary, AI shows great potential in a mobile network environment, where different network challenges can be dealt with by jointly optimizing communication, computing, caching with intelligent control (e.g., DRL). Indeed, the DRL approach achieves breakthroughs due to its ability to make optimal resource allocation decisions, especially in a highly time-varying dynamic systems. The performance of the DRL approach in the 4C optimization can be demonstrated by using TensorFlow in computer simulations. TensorFlow is an ML system operating in large-scale and effectively applied to wireless Het-Nets. Specifically, TensorFlow utilizes unified dataflow graphs to address the computations in algorithms, states, and actions. It can map the dataflow graph nodes across several machines in a group and also in a machine across various computing devices, such as general-purpose GPUs, multi-core CPUs, and custom designed ASICs (aka Tensor Processing Units). Being an open source, TensorFlow can provide developers with flexibility and enable them to conduct experiments with innovative optimizations and training algorithms according to prior parameter server designs. Therefore, it supports diverse applications needing training and inference algorithms on advanced deep neural networks; see [108, 124, 125].

5.4 Classification of the Integration Approaches

Previously, several studies integrated key components/parts of 4C. The contributions of such efforts improved wireless networks performance and pave the way for the emergence and thriving of the i4C. This subsection dedicates itself to these essential concepts, providing good insights into the prospects of converg-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

Table 3. *Summary of existing efforts on the integration of communication and computing.*

Integration	Network	Main Contribution	Objective			Evaluation Method	Reference
			Latency Minimization	Energy Saving	Other Performance Metrics		
Communication and Computing	Wireless Het-Nets	Formulating a joint communication and computing resources optimization problem	√			Numerical	[66]
	Mobile wireless network	Proposing a joint communication and computation resource allocation for a TDMA-based multi-user MECO system	√			Numerical	[109]
	Mobile wireless network	Proposing a joint communication and computation cooperation approach	√	√		Simulation	[110]
	Mobile wireless network	Proposing a joint communication and computation cooperation approach		√		Numerical and simulation	[73]
	Wireless Het-Nets	Presenting an integrated framework for computation offloading and resource allocation in MEC		√		Numerical and simulation	[111]
	Mobile wireless network	Hybrid pre-coding with communication and computational capabilities algorithm		√		Simulation	[112]
	Wireless Het-Nets	Formulating a joint optimization for transmission and processing delays			√	Simulation	[113]
	Wireless Het-Nets	Integrated framework of VFC for communication and computing resources			√	Simulation	[114]
	Wireless Het-Nets	Proposing an MEC collaborative architecture for resource sharing among MEC-BSs in UDN			√	Simulation	[115]

ing 4C in 5G, 6G, and beyond networks. Therefore, we focus on different approaches to i4C, classifying various works that converged or explored these four underlying functionalities.

Specifically, we classify the existing studies pertaining to i4C into integration of: 1) communication, computing, and caching; 2) communication, computing, and control; 3) communication and computing; 4) communication and caching; 5) communication and control; 6) computing and caching; 7) computing and control; and 8) caching and control. However, due to limited space, we summarize the contributions of some of these efforts based their objectives (i.e., time saving, energy saving, and other metrics) in Tables 3, 4, 5, and 6. Thus, this survey does not cover all existing literature related to i4C.

In particular, the focus of discussion in Table 3 lies in the convergence of communication and computing, where the major thrust targeted low latency, energy minimization, and other key metrics. Table 4 presents the summary of existing studies on the integration of communication and control with focus on realizing low latency and other metrics. The themes of the efforts summarized in Table 5 lies in coupling computing and caching with the aim of realizing low latency, low energy consumption, latency and energy savings, and other performance metrics. Table 6 summarizes some existing efforts devoted to integrating communication, computing, and caching with the goals of attaining ultra-low latency, low energy consumption, low latency and energy savings, and other performance metrics.

5.5 Integration of Sensing and Communication (ISAC)

In 6G wireless networks, various intensive computing services are expected to pop up. Distributed computing plays the role of a key 6G enabler that collectively utilizes pervasive sensing, communication, and computing functionalities in UEs, MEC/edge servers, and network nodes. With distributed com-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

Table 4. *Summary of existing efforts on the integration of communication and control.*

Integration	Network	Main Contribution	Objective		Evaluation Method	Reference
			Latency Minimization	Other Performance Metrics		
Communication and Control	V2V wireless networks	Proposing an integrated control system and V2V wireless communication co-design framework	√		Simulation	[26]
	V2V wireless networks	Proposing an integrated control system and V2V wireless communication co-design framework	√		Simulation	[116]
	Wireless vehicular networks	Applying a Smith predictor	√		Experimental	[117]
	Wireless networks	Joint control and communication system design	√		Simulation	[118]
	Wireless NCS	Proposing a joint control and communication design		√	Simulation	[119]
	Wireless NCS	Proposing three self-triggered control strategies		√	Numerical	[120]
	Wireless network	Joint UAV trajectory and power control scheme		√	Numerical	[121]
	Wireless cellular networks	Proposing an adaptive distributed power control algorithm		√	Simulation	[122]
	Intervehicle communication networks	Proposing a novel adaptive switched control algorithm		√	Simulation	[123]

Table 5. *Summary of existing efforts on the integration of computing and caching.*

Integration	Network	Main Contribution	Objective			Evaluation Method	Reference
			Latency Minimization	Energy Saving	Other Performance Metrics		
Computing and Caching	Wireless Het-Nets	Proposing optimal offloading with caching enhancement scheme (OOCS)	√			Simulation	[126]
	Wireless Het-Nets	Proposing a data allocation algorithm and an offloading scheduling algorithm	√			Simulation	[127]
	Wireless Het-Nets	Formulating an integrated model of computation offloading, caching, and resource allocation	√			Simulation	[128]
	Wireless Het-Nets	Task caching and computation offloading scheme	√	√		Simulation	[129]
	Wireless Het-Nets	Proposing OREO to jointly optimize dynamic service caching and computation offloading	√	√		Simulation	[130]
	Wireless Het-Nets	Constructing a joint computation offloading scheduling and caching scheme	√	√		Simulation	[131]
	Wireless Het-Nets	Computation and caching resources allocation in MEC			√	Simulation	[32]
	Wireless Het-Nets	Formulating a joint collaborative caching and processing problem as an ILP			√	Simulation	[132]
	Wireless Het-Nets	Proposing a joint collaborative caching and processing framework			√	Simulation	[133]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

Table 6. *Summary of existing efforts on the integration of communication, computing, and caching.*

Integration	Network	Main Contribution	Objective			Evaluation Method	Reference
			Latency Minimization	Energy Saving	Other Performance Metrics		
Communication, Computing, and Caching	Edge -cloud	Proposing a joint optimization of communication, computing, and caching on edge cloud, called Edge-CoCaCo	√			Simulation	[134]
	Mobile vehicular Networks	Devising a joint communication, computing, and caching system model	√			Simulation	[135]
	Virtualized Network	Proposing an Air-ground integrated MEC architecture	√			Numerical and simulation	[136]
	Wireless Het-Nets	Proposing Hybrid IoT to enable efficient transmission, computing, and caching of big data	√	√		Numerical and simulation	[137]
	Mobile wireless network	Developing implementation framework for mobile VR delivery	√	√		Numerical	[138]
	Mobile wireless network	Presenting a novel MEC-based mobile VR delivery framework	√	√		Numerical	[139]
	F-RANs	Enabling a joint caching and computing policy using the communication, computing, and caching resource allocation problem	√	√		Numerical	[140]
	Multiuser MEC-based wireless network	Formulating joint optimization of computing and caching policy			√	Analytical/ Numerical	[141]
	Wireless Het-Nets	Formulating a joint computation offloading, spectrum resource and computation resource allocation, and content caching optimization			√	Simulation	[96]
	Wireless Het-Nets	Formulating a joint computation offloading, resource allocation, and content caching optimization			√	Simulation	[142]
	Virtualized Het-Nets	Designing a novel information -centric Het-Nets framework			√	Simulation	[39]
	Virtualized Het-Nets	Designing a novel information -centric Het-Nets framework for sharing communication, computing, and caching resources			√	Simulation	[143]

puting, computation-intensive tasks can be partitioned into several subtasks and allocated to various network nodes for parallel collaborative computing. For instance, wireless distributed learning and reasoning could be employed to intelligently forecast future traffic pattern, network bottleneck, and resource availability based on archived and real-time big data in mobile networks[144]. Such data can be complex and enormous and have pivotal roles to play in combining sensing and communication. Data are exchanged among numerous UEs and also between the edge/-cloud servers and UEs, including connected IoT devices, IoV, autonomous vehicles. Various built-in sensors in such devices collect the data for swift processing and analysis. Therefore, while moving toward 6G, we will be witnessing sensor data sharing (exchange of information about the environment) among autonomous vehicles and edge/cloud servers.

The ability to interact by sharing data (information) among various nodes and continuously sense the dynamically varying conditions of the environment is one of the core drivers for vehicular (autonomous) networks. To realize the autonomous systems, different functionalities/subsystems need to converge. In the 6G era, communication and sensing will be tightly fused together to support multiple autonomous systems, e.g., UAVs, autonomous vehicles, and Industry 4.0 [23]. To this end, Wild et al. [145] focused on join communications and sensing (JCAS) design aspects for beyond 5G and 6G networks. The authors offered key drivers for integrating communications and sensing (e.g., AI and signal processing, massive bandwidth, denser networks, and MIMO and beamforming), analyzed the waveform appropriate for communications and radar sensing, considered various techniques for converging communications and sensing capabilities, discussed visions for advanced communications and sensing systems built upon distributed MIMO, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

presented many research challenges for JCAS, which should be overcome to attain natively integrated communications and sensing in the 6G mobile networks.

Beyond ISAC or JCAS, 6G is expected to be the 1st generation of cellular networks where localization, sensing, communication, and computing functionalities will be tightly integrated. One of the global 6G initiatives is the Hexa-X flagship project that consolidates 25 major participants from academia and industry; among the explicit objectives of the Hexa-X project is researching fundamentally new RATs, high-resolution locations, and sensing. The focus of 6G is to merge physical, digital, and human worlds, and the bridge that connects these worlds is the capability to sense, localize, and track physical objects. Under Hexa-X lie the integral parts of 6G comprising vision, radar, localization, and sensing. This will offer the ultra-high performance required for supporting location accuracies and latencies foreseen in the recognized/identified use-case families; it will also result in the tight integration of communications, radar, computing, localization, and sensing at both physical and software levels [146].

Recently, the integration of communication, computing, caching, control, and sensing gains considerable interest from academia and industry due to its strong potential in 6G networks. Chowdhury et al. [23] described that 5G largely overlooks the integration of sensing, communication, computing, intelligence, and control functionalities. 6G promises to fulfill this lagging and cope with the 5G constraints for accommodating new challenges. In particular, the requirements of emerging IoIT/IoE applications, including XR, haptics, telemedicine, automation, and robotics, will surpass the capabilities of 5G and necessitate the integration of communication, computing, caching, control, and sensing in the 6G and beyond networks [23, 147]. In other words, converging AR and VR cannot suffice the challenging requirements of several applications, such as near-real-person video conferencing, remote surgery/diagnosis, and ultra-high-resolution remote sensing for remote exploration. Today, holographic teleportation is considered as a potential replacement for AR/VR-enabled solutions[148]. The holographic and high-precision communication for haptics and Tactile applications will be supported by the two fundamental drivers for 6G, viz., IoE and mobile Internet, to realize comprehensive sensory responses, i.e., hearing, vision, smell, touch, and taste. This calls for processing enormous data in near-real time, ultra-high throughput (about 1-5 Tbps), and ultra-low latency (about 1 ms). Hence, there is need for integrating sensing, control, communication, caching, and computing capabilities in the 6G networks. Based on intelligent control, the integration of these capabilities will enable the networks to optimally decide what objects need to be sensed, what computational tasks need to be processed by which computing resources, and what data need to be stored by which caching resources.

However, realizing the integration of communication, computing, caching, control, and sensing encounters some difficulties due to: i) complex communication resources (e.g., multi-dimensional radio and x-haul resources), ii) multi-layered computing resources (e.g., x-computing), iii) multi-layered caching resources, and iv) a large quantity of sensing objects (e.g., environments, humans, and things) for verticals and IoE applications. Fortunately, AI emerges capable of choosing appropriate sensing objects and competently managing communication, computing, and caching resources through learning from data, training, predicting, and decision making [48]. Thus, incorporating AI into wireless network domains will efficiently overcome these challenges to achieve the integration of communication, computing, caching, control, and sensing.

VI. OPEN CHALLENGES AND FUTURE DIRECTIONS

In the previous section, we present a plethora of research works pertaining to the i4C. In this section, we will introduce: A) a number of open challenges and B) future directions for potential efforts on the i4C.

6.1 Open Challenges

Of course, the i4C can bring promising opportunities for future networks. Despite this bright future, the i4C raises several challenges that call for proper handling prior to fully implement it in 6G and beyond networks. To this end, this subsection highlights some critical challenges pertaining to the i4C.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

6.1.1 The 4C Tradeoff

In [46], it was proved that with an integrated system harnessing synergistic combinations of different functionalities/resources, the same types of services can be realized. Nevertheless, the tradeoffs amongst the intrinsic 4C functionalities/resources for each type of service have to be fully investigated separately against the associated performance metrics and the constraints of the functionalities/resources. Hence, the optimal tradeoff of the functionalities/resources for each separate case, which is obviously important to determine, remains a key challenge in this context; see [13].

6.1.2 Mobility in 4C

One of the key factors that may account for frequent channel disconnection between the mobile UEs and the network edges is mobility. Due to the dynamic characteristics of wireless network parameters, including jitter, bandwidth, latency, and so on, the QoS of an application can be deteriorated when a mobile UE happens to be in moving state [31].

Wang et al. [14] pointed out that user mobility seriously disrupts both caching and computation offloading decisions and results in recurring handovers among the servers of the network edges. In other words, the user mobility and the short coverage of network edges contribute immensely to: i) degrading the efficiency of wireless networks, ii) drastic reduction in users' QoS, and iii) interrupting ongoing edge services [149]. In short, frequent user mobility can severely limit the performance of the i4C, since joint consideration of communication channel and computation capacity is essential for offloading the computational tasks. To that end, mobility management becomes necessary and needs to be redesigned in 4C scenarios [150].

6.1.3 Interference in 4C

Signal interference plays a critical role in the wireless communication channel. To be more specific, if there is interference between different UEs, the control signal may be lost, thereby giving rise to further problems, such as high energy consumption, transmission delay, and lower bandwidth. This implies an adverse consequences on the communication resource, on which the other three key resources may depend. Hence, the wireless channel interference remains a key challenge that requires further research investigations.

6.1.4 Security Issues in 4C

Today, 4C functionalities are mostly hosted at the close proximity of UEs (e.g., the network edges) to save both energy and time. One of the finest examples of such hosts are the MEC servers, and deploying such servers to the network edges amounts to exposing them to vulnerable security threats. Moreover, disrupting the severs by any physical or cyber threat is tantamount to disrupting the 4C services/functionalities, limiting the performance of the integrated system of 4C. Therefore, securing the 4C functionalities/services at the network edges becomes a technical challenge to be addressed.

6.1.5 Data Privacy

Data privacy is another critical challenge that quests for further investigations in the 4C scenarios. In the integrated system of 4C, substantial amount of data may be offloaded from UEs to the servers of the network edges for real-time analytics, processing, and caching services. Such data could be: i) metadata, such as geographical locations, timestamps, and so on; ii) computing strategies, which includes computation offloading strategy; and iii) monitoring data [151].

Indeed, the emergence of edge computing contributes to dramatic reduction of end-to-end latency; however, data privacy will more than likely face severe challenges. The threat of data tampering and information leakage is aggravated as a result of wireless channel properties for computation offloading. For instance, if an edge server happens to be under the control of a malicious eavesdropper, the enterprise multimedia data may be transmitted to the server under the control. In this respect, the multimedia data can be handily eavesdropped or even tempered by the eavesdropper. Presently, data encryption is employed to guarantee data privacy and security. Nevertheless, due to the wireless channel characteristics and high complexity in computation, which result in latency and low QoE, data encryption may not sufficiently address the issues of data privacy in 4C scenarios [152].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

6.1.6 Ultra-low Latency Requirements in 4C:

Convergence of 4C has broad range of mission-critical applications, among which are the UAV flight control systems, Tactile Internet, XR applications, IoIT, IoE, autonomous vehicles, and Internet of vehicles. Such applications usually require extremely low latency in a few tens of milliseconds. Unfortunately, the conventional wireless systems cannot meet such low latency requirements, which in return create additional intense challenges for the integrated system of 4C; see [13].

6.2 Future Directions

6.2.1 5G, 6G, and Beyond Networks

Nowadays, myriad intelligent applications and use cases, such as XR, IoE, IoIT, autonomous vehicles, blockchain, Tactile Internet, Telesurgery, et cetera, surface with different service requirements. 5G, having 4C functionalities interacting at the proximity of such applications, promises to handle their diverse QoS requirements. However, user devices and their intelligent applications keep proliferating exponentially with stringent requirements, driving the 5G capabilities to their limits. This necessitates researchers in academia and industry to look beyond the boundaries of 5G networks. To fully support the emerging intelligent applications and use cases in the coming decades, there is need to look forward to converging the 4C functionalities in 6G and beyond networks.

6.2.2 Integration

The recent emergence of edge intelligence will undoubtedly trigger further research investigation. The point is that various network functionalities/resources, including communication, computing, caching, and control, are involved in the architecture of the edge networks. Yet, a systematic convergence of 4C (with the capability of realizing the system-level (optimal) performance) is far from being concluded [14]. Besides, there is still an urgent need for more holistic and intelligent control schemes to optimally control, coordinate, and integrate the network edge functionalities. Thus, more mechanisms for the i4C functionalities are required at the proximity of UEs. On the other hand, several emerging applications and use cases necessitate the convergence of communication, computing, control, and sensing in the 6G era. Suffice it to say, research on integration has to continue in the future.

6.2.3 Fundamental Relations behind 4C

Most of the existing research efforts conducted on the i4C focused on improving capacity, latency minimization, and energy savings in mobile networks. Of course, there are many more promising solutions behind a fully integrated system of 4C. However, such system can be successfully realized by leveraging the full synergy, capabilities, and tradeoff of the 4C functionalities. Jiao et al. [153] argued that the fundamental benefits behind utilizing the caching and computing functionalities/resources in mobile networks have not yet been studied effectively. Another effort in [6] pointed out that the ultimate synergy behind a fully integrated solution of 4C is not nearly well understood. To this end, synergistic collaboration, tradeoffs, and capabilities of 4C need to be further studied in order to fully leverage and benefit from the integrated solution of 4C in 6G and beyond networks.

6.2.4 The Capacity of 4C

Although communication capacity can be determined in terms of Shannon information transmission theorem, the theoretical capacity of each measure (dimension) of caching, computing, and control functionalities is not yet determined. Hence, determining the theoretical capacity of each individual functionality of 4C remains a potential direction that quest for further investigations. The classical information theoretical model (by which instantaneous rate regions is addressed) is not directly applicable to the converged system of these functionalities. This is because it lacks ability to efficiently address caching-induced non-causality in the system [13], [46]. To sum up, an effort in [154] uncovered the insufficiency of the conventional concept of rate capacity to portray the network strength (ability) to deliver (release) non-private contents for several UEs. The effort, which aimed at measuring the impact of caching in content releasing, introduced a so-called content rate for measuring the rate at which the amount of cached data is released to UEs through a shared wireless channel.

The explicit role that computing plays in the integrated system capacity measurement and calculation is not properly understood, despite numerous existing in-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.23919/JCC.2023.00.039, China Communications

vestigations devoted to integration of wireless communications and computing. A typical approach that has to do with this is network coding, which distinguishes algebraic operation (computing) from communication operation. Other approaches, such as collaborative transmission [155] and distributed MIMO [156], do not decouple computing and communication, as logical operations and channel coding across information streams are entwined. Obviously, a bound together capacity analysis that portrays communication, computing, and caching resources in canonical frame, bears considerable hypothetical value, and as such, is in critical interest; see [13, 46].

6.2.5 Real-Time Decision Making

Due to their proximity to UEs, MEC and other edge networks can track their real-time information like user's location, behavior, and resources environment [14],[7]. Delivering context-aware services to UEs can be enabled by inference based on such information. For example, for video guidance in a museum, the subscribers' interests can be predicted/learnt through an AR application, based on their (subscribers') positions in the museum, for delivering contents, such as artworks and antiques. The CTrack system is another example, which tracks and predicts multiple subscribers' trajectories by using BS fingerprints for routing, navigation, monitoring, and personalized trip management [7]. The success behind realizing such real-time control systems requires deep insights into the communications, computing control [157] and caching theories. On top of that, proactive resource allocation requires the use of different levels of real-time information, i.e., application, network, and UE levels [14].

6.2.6 Intelligence for the i4C

Recent research investigations resort to intelligence to overcome critical issues of i4C. This may not be unconnected with the complexity of the integrated system and specific use cases in MEC. The study in [44] raised some key points behind considering intelligence for resource optimization/integration. As discussed in the sequel, AI frontiers, such as ML, DRL and FL, will have pivotal roles to play in 4C scenarios.

Machine Learning: The 5G/6G networks will be endowed with the swarm intelligence and node intelligence to improve their efficiency. Against conventional objective function of a single component, a trade-off bounded by various factors, such as energy consumption, delay, capacity, complexity, and so on, will be dealt with for management and allocation of resources [158]. Hence, heterogeneous network devices, different QoS demands, as well as large state and action spaces will greatly complicate the i4C in 5G, 6G, and beyond networks. In the face of such complication, future wireless networks may rely on ML for online and/or fully-distributed algorithms. Moreover, ML has capability of dealing with the challenges of closed-form solution, problem formulation, and other issues of channel modeling inflicted by model-free wireless networks [8]. In fact, by leveraging learning based on try and error experiments, ML is found efficiently capable of handling multi-objective optimization problems, especially in light of managing the multi-agent collaborative networks [158].

Needless to say, DRL will potentially play major roles in breaking the complexity of join optimization of 4C especially in extremely time-varying MEC systems. DRL emerges through the coupling of reinforcement learning algorithm with deep learning to combat considerable input data amount and determine the optimal policy for the complicated resource allocation problems. In DRL, deep Q-network (DQN) can be leveraged for approximating the Q value function [13]. In [105, 106], Q learning and DQN have been studied to overcome some challenges related to joint optimization of 4C; see [8]. Likewise, [61, 107, 159] applied DRL approach to investigate the integration of communication, computing, and caching.

Federated Learning: Data privacy is a critical challenge that calls for more attention. In contrast to conventional approaches of ML, which have no room for preserving the privacy of training data, FL can enable UEs to collaboratively learn a shared model while locally keeping their respective data. In other words, FL can allow the distribution of training data across individual UEs. Hence, the limitations of distributed learning can be dealt with. Such limitations include: i) low efficiency caused by heterogeneous capabilities in UEs and network states, ii) insufficient time and training data, iii) lopsidedness in the number of the training data samples, as well as iv) non-independent and identically distributed data between the UEs.

In a computation offloading technique, a significant number of UEs may offload their tasks for remote

computation and caching. In tradition, the UEs decide whether to offload the tasks or not by reporting their individual information, which involves battery lifespan, channel gain, computing capabilities, and so on, to the network edge. Unfortunately, malicious eavesdroppers can access and even use the information illegally to obtain the locations of the UEs. Applying FL will allow each individual UE to download the master model from the network edge and thereby learn the computation offloading decisions based on its local information only. Moreover, based on the updates obtained from the UEs, the network edge will be responsible for the master model update. In this way, FL will preserve the privacy of data and bring distributed offloading decisions. In summary, FL can be considered as one of the finest potential solutions to resource allocation/integration issues. Thus, it is foreseen to serve as a sharp tool for various challenges related to resource allocation optimization in MEC [8].

VII. CONCLUSION

This article brings an extensive survey on the i4C, which becomes indispensable due to recent growth of smart devices and their emerging mission-critical applications. The survey starts with providing a snapshot of different aspects of the i4C, including motivations, some potential applications and use cases, and key enabling technologies. To lay the foundation of the integration in 5G, 6G, and beyond networks, the article offers a firsthand tutorial on various models of 4C. Then, it reviews several state-of-the-art efforts on the i4C, placing emphasis on recent trends of conventional optimization and AI-based integration approaches. It disscuses the convengence of communication and sensing and classifies different approaches of resource integration. Finally, open challenges and future directions are discussed.

References

[1] P. Fan, "Coping with the big data: Convergence of communications, computing and storage," *China Communications*, vol. 13, no. 9, pp. 203–207, 2016.

[2] K. Wang and K. Yang, "Power-minimization computing resource allocation in mobile cloud-radio access network," in *IEEE International Conference on Computer and Information Technology (CIT)*. IEEE, 2016, pp. 667–672.

[3] H. Zhang, G. Liu *et al.*, "Dynamics of communication, caching and computing resource sharing: A game model," in *IEEE 86th Vehicular Technology Conference (VTC-Fall)*. IEEE, 2017, pp. 1–5.

[4] V. Lima, M. Eisen *et al.*, "Model-free design of control systems over wireless fading channels," *arXiv preprint arXiv:2009.01751*, 2020.

[5] J. Zeng, J. Sun *et al.*, "Mobile edge communications, computing, and caching (mec3) technology in the maritime communication network," *China Communications*, vol. 17, no. 5, pp. 223–234, 2020.

[6] S. Andreev, O. Galinina *et al.*, "Exploring synergy between communications, caching, and computing in 5g-grade deployments," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 60–69, 2016.

[7] Y. Mao, C. You *et al.*, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[8] Q.-V. Pham, F. Fang *et al.*, "A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116 974–117 017, 2020.

[9] L. Bariah, L. Mohjazi *et al.*, "A prospective look: Key enabling technologies, applications and open research topics in 6g networks," *IEEE Access*, vol. 8, pp. 174 792–174 820, 2020.

[10] H. Yang, A. Alphones *et al.*, "Artificial-intelligence-enabled intelligent 6g networks," *IEEE Network*, vol. 34, no. 6, pp. 272–280, 2020.

[11] K. B. Letaief, W. Chen *et al.*, "The roadmap to 6g: Ai empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, 2019.

[12] D. Xu, T. Li *et al.*, "Edge intelligence: Architectures, challenges, and applications," *arXiv preprint arXiv:2003.12172*, 2020.

[13] C. Wang, Y. He *et al.*, "Integration of networking, caching, and computing in wireless systems: A survey, some research issues, and challenges," *IEEE Communications Surveys & Tu-*

torials, vol. 20, no. 1, pp. 7–38, 2017.

[14] S. Wang, X. Zhang *et al.*, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.

[15] M. A. Bouras, F. Farha *et al.*, "Convergence of computing, communication, and caching in internet of things," *Intelligent and Converged Networks*, vol. 1, no. 1, pp. 18–36, 2020.

[16] K.-D. Kim and P. R. Kumar, "Cyber–physical systems: A perspective at the centennial," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1287–1308, 2012.

[17] U. Challita, H. Ryden *et al.*, "When machine learning meets wireless cellular networks: Deployment, challenges, and applications," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 12–18, 2020.

[18] B. Zhang, "A collective communication layer for the software stack of big data analytics," in *IEEE International Conference on Cloud Engineering Workshop (IC2EW)*. IEEE, 2016, pp. 204–206.

[19] A. Ndikumana, N. H. Tran *et al.*, "Joint communication, computation, caching, and control in big data multi-access edge computing," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1359–1374, 2019.

[20] M. Torabzadehkashi, S. Rezaei *et al.*, "Catalina: in-storage processing acceleration for scalable big data analytics," in *27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. IEEE, 2019, pp. 430–437.

[21] P. Zhang, K. Yu *et al.*, "Quantcloud: big data infrastructure for quantitative finance on the cloud," *IEEE Transactions on Big Data*, vol. 4, no. 3, pp. 368–380, 2017.

[22] A. I. Maarala, M. Rautiainen *et al.*, "Low latency analytics for streaming traffic data with apache spark," in *IEEE International Conference on Big Data (Big Data)*. IEEE, 2015, pp. 2855–2858.

[23] M. Z. Chowdhury, M. Shahjalal *et al.*, "6g wireless communication systems: Applications, requirements, technologies, challenges, and research directions," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 957–975, 2020.

[24] A. Jahid, M. H. Alsharif *et al.*, "The convergence of blockchain, iot and 6g: Potential, opportunities, challenges and research roadmap," *arXiv preprint arXiv:2109.03184*, 2021.

[25] W. Jiang, B. Han *et al.*, "The road towards 6g: A comprehensive survey," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 334–366, 2021.

[26] T. Zeng, O. Semiari *et al.*, "Joint communication and control for wireless autonomous vehicular platoon systems," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7907–7922, 2019.

[27] S. Sukhmani, M. Sadeghi *et al.*, "Edge caching and computing in 5g for mobile ar/vr and tactile internet," *IEEE MultiMedia*, vol. 26, no. 1, pp. 21–30, 2018.

[28] B. Wang, Y. Sun *et al.*, "Hierarchical matching with peer effect for low-latency and high-reliable caching in social iot," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1193–1209, 2018.

[29] M. A. Bouras, A. Ullah *et al.*, "Synergy between communication, computing, and caching for smart sensing in internet of things," *Procedia computer science*, vol. 147, pp. 504–511, 2019.

[30] G. Gao, Y. Wen *et al.*, "vcache: Supporting cost-efficient adaptive bitrate streaming," *IEEE MultiMedia*, vol. 24, no. 3, pp. 19–27, 2017.

[31] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," *10th International Conference on Intelligent Systems and Control (ISCO)*, pp. 1–8, 2016.

[32] A. Ndikumana, S. Ullah *et al.*, "Collaborative cache allocation and computation offloading in mobile edge computing," in *2017 19th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, 2017, pp. 366–369.

[33] Y. Tao, C. You *et al.*, "Stochastic control of computation offloading to a helper with a dynamically loaded cpu," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1247–1262, 2019.

[34] S. Kim, "5g network communication, caching, and computing algorithms based on the two-tier

game model," *Etri Journal*, vol. 40, no. 1, pp. 61–71, 2018.

[35] Q. Chen, F. R. Yu *et al.*, "Joint resource allocation for software-defined networking, caching, and computing," *IEEE/ACM Transactions on Networking*, vol. 26, no. 1, pp. 274–287, 2018.

[36] J. A. Cabrera, R.-S. Schmoll *et al.*, "Softwarization and network coding in the mobile edge cloud for the tactile internet," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 350–363, 2018.

[37] W. Zhuang, Q. Ye *et al.*, "Sdn/nfv-empowered future iov with enhanced communication, computing, and caching," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 274–291, 2019.

[38] Q.-Y. Zhang, X.-W. Wang *et al.*, "Software defined networking meets information centric networking: A survey," *IEEE Access*, vol. 6, pp. 39 547–39 563, 2018.

[39] Y. Zhou, F. R. Yu *et al.*, "Resource allocation for information-centric virtualized heterogeneous networks with in-network caching and mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 11 339–11 351, 2017.

[40] Y. Zhou, R. F. Yu *et al.*, "Communications, caching, and computing for next generation hetnets," *IEEE Wireless Communications*, vol. 25, no. 4, pp. 104–111, 2018.

[41] E. Peltonen, M. Bennis *et al.*, "6g white paper on edge intelligence," *arXiv preprint arXiv:2004.14850*, 2020.

[42] S. Deng, H. Zhao *et al.*, "Edge intelligence: The confluence of edge computing and artificial intelligence," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7457–7469, 2020.

[43] A. A. Corici, Y. C. Hu *et al.*, "Edge intelligence," *White Paper, IEC*, 2018.

[44] X. Wang, Y. Han *et al.*, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 869–904, 2020.

[45] S.-W. Ko, K. Han *et al.*, "Wireless networks for mobile edge computing: Spatial modeling and latency analysis," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5225–5240, 2018.

[46] H. Liu, Z. Chen *et al.*, "The three primary colors of mobile systems," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 15–21, 2016.

[47] J. Baumgarten and T. Kuerner, "Lte downlink link-level abstraction for system-level simulations," in *European Wireless 2014; 20th European Wireless Conference*. VDE, 2014, pp. 1–5.

[48] Z. Zhang, Y. Xiao *et al.*, "6g wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, 2019.

[49] Z. Tan, F. R. Yu *et al.*, "Virtual resource allocation for heterogeneous services in full duplex-enabled scns with mobile edge computing and caching," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1794–1808, 2017.

[50] N. Saha and R. Vesilo, "An evolutionary game theory approach for joint offloading and interference management in a two-tier hetnet," *IEEE Access*, vol. 6, pp. 1807–1821, 2017.

[51] J. Zhang, S. Qu *et al.*, "Regularized interference alignment for heterogeneous networks," in *8th IEEE international conference on communication software and networks (ICCSN)*. IEEE, 2016, pp. 201–205.

[52] S. Mu, Z. Zhong *et al.*, "Latency constrained partial offloading and subcarrier allocations in small cell networks," in *ICC IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–7.

[53] A. S. Hamza, S. S. Khalifa *et al.*, "A survey on inter-cell interference coordination techniques in ofdma-based cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1642–1670, 2013.

[54] L. Sharnagat and H. Harichand, "Method of resource allocation in ofdma using convex optimization," in *Fifth International Conference on Communication Systems and Network Technologies*. IEEE, 2015, pp. 407–411.

[55] H.-C. Yang and M.-S. Alouini, "Characterizing energy efficiency of wireless transmission for green internet of things: A data-oriented approach," *arXiv preprint arXiv:1805.11725*, 2018.

[56] X. Deng and A. M. Haimovich, "Information rates of time varying rayleigh fading channels," in *IEEE International Conference on Commu-*

nications (IEEE Cat. No. 04CH37577), vol. 1. IEEE, 2004, pp. 573–577.

[57] J. Liu, N. Elia *et al.*, "Capacity-achieving feedback schemes for gaussian finite-state markov channels with channel state information," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 3632–3650, 2015.

[58] J. Wang, J. Cai *et al.*, "New channel model for wireless communications: Finite-state phase-type semi-markov channel model," in *IEEE International Conference on Communications*. IEEE, 2008, pp. 4461–4465.

[59] X. Wang, Y. Han *et al.*, "In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Network*, vol. 33, no. 5, pp. 156–165, 2019.

[60] Y. He, C. Liang *et al.*, "Integrated computing, caching, and communication for trust-based social networks: A big data drl approach," in *IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.

[61] Y. He, C. Liang, F. R. Yu, and Z. Han, "Trust-based social networks with computing, caching and communications: A deep reinforcement learning approach," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 66–79, 2018.

[62] Z. Luo, M. LiWang *et al.*, "Energy-efficient caching for mobile edge computing in 5g networks," *Applied sciences*, vol. 7, no. 6, p. 557, 2017.

[63] K. Cheng, Y. Teng *et al.*, "Energy-efficient joint offloading and wireless resource allocation strategy in multi-mec server systems," in *IEEE international conference on communications (ICC)*. IEEE, 2018, pp. 1–6.

[64] X. Chen, "Decentralized computation offloading game for mobile cloud computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 4, pp. 974–983, 2014.

[65] N. Abbas, Y. Zhang *et al.*, "Mobile edge computing: A survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2017.

[66] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5g heterogeneous networks," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 45–55, 2014.

[67] B. P. Rimal, D. P. Van, and M. Maier, "Mobile-edge computing vs. centralized cloud computing in fiber-wireless access networks," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2016, pp. 991–996.

[68] S. Singh, "Optimize cloud computations using edge computing," in *International Conference on Big Data, IoT and Data Science (BID)*. IEEE, 2017, pp. 49–53.

[69] S. Deshmukh and R. Shah, "Computation offloading frameworks in mobile cloud computing: a survey," in *IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*. IEEE, 2016, pp. 1–5.

[70] M. Othman, S. A. Madani *et al.*, "A survey of mobile cloud computing application models," *IEEE communications surveys & tutorials*, vol. 16, no. 1, pp. 393–413, 2013.

[71] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[72] M. Liu and Y. Liu, "Price-based distributed offloading for mobile-edge computing with computation capacity constraints," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 420–423, 2017.

[73] X. Cao, F. Wang *et al.*, "Joint computation and communication cooperation for energy-efficient mobile edge computing," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4188–4200, 2018.

[74] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3590–3605, 2016.

[75] Z. Sheng, C. Mahapatra *et al.*, "Energy efficient cooperative computing in mobile wireless sensor networks," *IEEE Transactions on Cloud Computing*, vol. 6, no. 1, pp. 114–126, 2015.

[76] X. Yang, Z. Chen *et al.*, "Communication-constrained mobile edge computing systems for wireless virtual reality: Scheduling and tradeoff," *IEEE Access*, vol. 6, pp. 16 665–16 677, 2018.

[77] J. Guo, Z. Song *et al.*, "Energy-efficient re-

source allocation for multi-user mobile edge computing," in *GLOBECOM IEEE Global Communications Conference*. IEEE, 2017, pp. 1–7.

[78] Y. Cui, W. He *et al.*, "Energy-efficient resource allocation for cache-assisted mobile edge computing," in *IEEE 42nd Conference on Local Computer Networks (LCN)*. IEEE, 2017, pp. 640–648.

[79] Y. Zhang, J. He, and S. Guo, "Energy-efficient dynamic task offloading for energy harvesting mobile cloud computing," in *IEEE international conference on networking, architecture and storage (NAS)*. IEEE, 2018, pp. 1–4.

[80] L. Chen, Z. Chen *et al.*, "Joint optimization of communications and computing resources allocation for deterministic transmission in wireless edge networks," vol. 19, no. 5. IEEE, 2022, pp. 1–11.

[81] X. Chen, L. Jiao *et al.*, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795–2808, 2015.

[82] G. S. Paschos, G. Iosifidis *et al.*, "The role of caching in future communication systems and networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1111–1125, 2018.

[83] S. Tarnoi, W. Kumwilaisak *et al.*, "Adaptive probabilistic caching technique for caching networks with dynamic content popularity," *Computer Communications*, vol. 139, pp. 1–15, 2019.

[84] V. Shivaram, N. Gupta, and K. S. Kamath, "Queuing models for different caching schemes by caching partial files," in *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018, pp. 1234–1238.

[85] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on information theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[86] M. A. Maddah Ali and U. Niesen, "Coding for caching: fundamental limits and practical challenges," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 23–29, 2016.

[87] S. Safavat, N. N. Sapavath, and D. B. Rawat, "Recent advances in mobile edge computing and content caching," *Digital Communications and Networks*, vol. 6, no. 2, pp. 189–194, 2020.

[88] Y. Tan, C. Han *et al.*, "Radio network-aware edge caching for video delivery in mec-enabled cellular networks," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2018, pp. 179–184.

[89] X. Wang, M. Chen *et al.*, "Cache in the air: Exploiting content caching and delivery techniques for 5g systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 131–139, 2014.

[90] E. Ramadan, P. Babaie, and Z.-L. Zhang, "Performance estimation and evaluation framework for caching policies in hierarchical caches," *Computer Communications*, vol. 144, pp. 44–56, 2019.

[91] V. Martina, M. Garetto, and E. Leonardi, "A unified approach to the performance analysis of caching systems," in *IEEE INFOCOM IEEE Conference on Computer Communications*. IEEE, 2014, pp. 2040–2048.

[92] S. Mehamel, K. Slimani *et al.*, "Energy-efficient hardware caching decision using fuzzy logic in mobile edge computing," in *6th International conference on future internet of things and cloud workshops (FiCloudW)*. IEEE, 2018, pp. 237–242.

[93] M. T. Beck, M. Werner *et al.*, "Mobile edge computing: A taxonomy," in *Proc. of the Sixth International Conference on Advances in Future Internet*. Citeseer, 2014, pp. 48–55.

[94] Y. Zhou, F. R. Yu *et al.*, "Information-centric wireless networks with mobile edge computing," *arXiv preprint arXiv:1706.09541*, 2017.

[95] J. Kwak, Y. Kim *et al.*, "Hybrid content caching in 5g wireless networks: Cloud versus edge caching," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3030–3045, 2018.

[96] C. Wang, C. Liang *et al.*, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 4924–4938, 2017.

[97] M. Farivar, X. Zho, and L. Chen, "Local voltage control in distribution systems: An incremental control algorithm," in *IEEE international con-*

ference on smart grid communications (SmartGridComm). IEEE, 2015, pp. 732–737.

[98] D. K. Molzahn, F. Dörfler *et al.*, "A survey of distributed optimization and control algorithms for electric power systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2941–2962, 2017.

[99] D. A. Novikov, "Hierarchical models in modern control theory," in *International Conference on Automation*. Springer, 2016, pp. 3–12.

[100] A. Abid, M. F. Manzoor *et al.*, "Challenges and issues of resource allocation techniques in cloud computing," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 14, no. 7, pp. 2815–2839, 2020.

[101] C. S. Magurawalage, K. Yang, and K. Wang, "Aqua computing: Coupling computing and communications," *arXiv preprint arXiv:1510.07250*, 2015.

[102] R. Huo, F. R. Yu *et al.*, "Software defined networking, caching, and computing for green wireless networks," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 185–193, 2016.

[103] Z. Zhou, X. Chen *et al.*, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[104] G. Zhu, D. Liu *et al.*, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE communications magazine*, vol. 58, no. 1, pp. 19–25, 2020.

[105] J. Li, H. Gao *et al.*, "Deep reinforcement learning based computation offloading and resource allocation for mec," in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2018, pp. 1–6.

[106] L. Hu, Y. Tian *et al.*, "Ready player one: Uav-clustering-based multi-task offloading for vehicular vr/ar gaming," *IEEE Network*, vol. 33, no. 3, pp. 42–48, 2019.

[107] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 44–55, 2017.

[108] L. T. Tan and R. Q. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 190–10 203, 2018.

[109] J. Ren, G. Yu *et al.*, "Latency optimization for resource allocation in mobile-edge computation offloading," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5506–5519, 2018.

[110] Y. Li, G. Xu *et al.*, "Communication and computation cooperation in wireless network for mobile edge computing," *IEEE Access*, vol. 7, pp. 106 260–106 274, 2019.

[111] J. Zhang, X. Hu *et al.*, "Energy-latency tradeoff for energy-aware offloading in mobile edge computing networks," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2633–2645, 2017.

[112] X. Ge, Y. Sun *et al.*, "Joint optimization of computation and communication power in multi-user massive mimo systems," *IEEE transactions on wireless communications*, vol. 17, no. 6, pp. 4051–4063, 2018.

[113] Y. Chen, E. Sun, and Y. Zhang, "Joint optimization of transmission and processing delay in fog computing access networks," in *9th International Conference on Advanced Infocomm Technology (ICAIT)*. IEEE, 2017, pp. 155–158.

[114] X. Hou, Y. Li *et al.*, "Vehicular fog computing: A viewpoint of vehicles as the infrastructures," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 6, pp. 3860–3873, 2016.

[115] T. Yang, H. Zhang *et al.*, "Computation collaboration in ultra dense network integrated with mobile edge computing," in *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2017, pp. 1–5.

[116] T. Zeng, O. Semiari *et al.*, "Integrated communications and control co-design for wireless vehicular platoon systems," in *IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.

[117] H. Xing, J. Ploeg, and H. Nijmeijer, "Smith predictor compensating for vehicle actuator delays in cooperative acc systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1106–1115, 2018.

[118] T. Zeng, M. Mozaffari *et al.*, "Wireless com-

munications and control for swarms of cellular-connected uavs," in *52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 719–723.

[119] A. Chamaken and L. Litz, "Joint design of control and communication in wireless networked control systems: A case study," in *Proceedings of the 2010 American Control Conference*. IEEE, 2010, pp. 1835–1840.

[120] S. Akashi, H. Ishii, and A. Cetinkaya, "Self-triggered control with tradeoffs in communication and computation," *Automatica*, vol. 94, pp. 373–380, 2018.

[121] Y. Huang, J. Xu *et al.*, "Cognitive uav communication via joint trajectory and power control," in *IEEE 19th international workshop on signal processing advances in wireless communications (SPAWC)*. IEEE, 2018, pp. 1–5.

[122] Y. Payasi, A. Shrivastava, and A. Jain, "Negotiation based adaptive distributed power control algorithm in wireless communication system," in *International Conference on Computer, Communication and Control (IC4)*. IEEE, 2015, pp. 1–5.

[123] Y. Abou Harfouch, S. Yuan, and S. Baldi, "An adaptive switched control approach to heterogeneous platooning with intervehicle communication losses," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1434–1444, 2017.

[124] M. Abadi, P. Barham *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[125] M. Abadi, A. Agarwal *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[126] S. Yu, R. Langar *et al.*, "Computation offloading with data caching enhancement for mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 11 098–11 112, 2018.

[127] W. Fan, Y. Liu *et al.*, "Terminalbooster: Collaborative computation offloading and data caching via smart basestations," *IEEE Wireless Communications Letters*, vol. 5, no. 6, pp. 612–615, 2016.

[128] J. Zhang, X. Hu *et al.*, "Joint resource allocation for latency-sensitive services over mobile edge computing networks with caching," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4283–4294, 2018.

[129] Y. Hao, M. Chen *et al.*, "Energy efficient task caching and offloading for mobile edge computing," *IEEE Access*, vol. 6, pp. 11 365–11 373, 2018.

[130] J. Xu, L. Chen, and P. Zhou, "Joint service caching and task offloading for mobile edge computing in dense networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 207–215.

[131] M. Liu, F. R. Yu *et al.*, "Computation offloading and content caching in wireless blockchain networks with mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 11 008–11 021, 2018.

[132] T. X. Tran, P. Pandey *et al.*, "Collaborative multi-bitrate video caching and processing in mobile-edge computing networks," in *13th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*. IEEE, 2017, pp. 165–172.

[133] T. X. Tran and D. Pompili, "Adaptive bitrate video caching and processing in mobile-edge computing networks," *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, pp. 1965–1978, 2018.

[134] M. Chen, Y. Hao *et al.*, "Edge-cocaco: Toward joint optimization of computation, caching, and communication on edge cloud," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 21–27, 2018.

[135] S. A. Kazmi, T. N. Dang *et al.*, "Infotainment enabled smart cars: A joint communication, caching, and computation approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8408–8420, 2019.

[136] Z. Zhou, J. Feng *et al.*, "An air-ground integration approach for mobile edge computing in iot," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 40–47, 2018.

[137] L. P. Qian, Y. Wu *et al.*, "Hybridiot: Integration of hierarchical multiple access and computation offloading for iot-based smart cities," *IEEE net-*

work, vol. 33, no. 2, pp. 6–13, 2019.

[138] Y. Sun, C. Zhiyon *et al.*, "Communication, computing and caching for mobile vr delivery: Modeling and trade-off," in *IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.

[139] Y. Sun, Z. Chen *et al.*, "Communications, caching, and computing for mobile virtual reality: Modeling and tradeoff," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7573–7586, 2019.

[140] T. Dang and M. Peng, "Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 7, pp. 1594–1607, 2019.

[141] Y. Sun, Z. Chen *et al.*, "Bandwidth gain from mobile edge computing and caching in wireless multicast systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3992–4007, 2020.

[142] C. Wang, C. Liang *et al.*, "Joint computation offloading, resource allocation and content caching in cellular networks with mobile edge computing," in *IEEE international conference on communications (ICC)*. IEEE, 2017, pp. 1–6.

[143] Y. Zhou, F. R. Yu *et al.*, "Virtual resource allocation for information-centric heterogeneous networks with mobile edge computing," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2017, pp. 235–240.

[144] Y. Zhou, L. Liu *et al.*, "Service-aware 6g: An intelligent and open network based on the convergence of communication, computing and caching," *Digital Communications and Networks*, vol. 6, no. 3, pp. 253–260, 2020.

[145] T. Wild, V. Braun *et al.*, "Joint design of communication and sensing for beyond 5g and 6g systems," *IEEE Access*, vol. 9, pp. 30 845–30 857, 2021.

[146] H. Wymeersch, D. Shrestha *et al.*, "Integration of communication and sensing in 6g: a joint industrial and academic perspective," in *IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2021, pp. 1–7.

[147] W. Saad, M. Bennis *et al.*, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, 2019.

[148] I. F. Akyildiz, A. Kak *et al.*, "6g and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133 995–134 030, 2020.

[149] S. Wang, J. Xu *et al.*, "A survey on service migration in mobile edge computing," *IEEE Access*, vol. 6, pp. 23 511–23 528, 2018.

[150] J. Wang, K. Liu *et al.*, "Learning based mobility management under uncertainties for mobile edge computing," in *IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.

[151] N. Hassan, K.-L. A. Yau, and C. Wu, "Edge computing in 5g: A review," *IEEE Access*, vol. 7, pp. 127 276–127 289, 2019.

[152] H. Nie, X. Jiang *et al.*, "Data security over wireless transmission for enterprise multimedia security with fountain codes," *Multimedia tools and applications*, vol. 79, no. 15, pp. 10 781–10 803, 2020.

[153] J. Jiao, X. Hong, and J. Shi, "Proactive content delivery for vehicles over cellular networks: The fundamental benefits of computing and caching," *China Communications*, vol. 15, no. 7, pp. 88–97, 2018.

[154] H. Liu, Z. Chen *et al.*, "On content-centric wireless delivery networks," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 118–125, 2014.

[155] D. Gesbert, S. Hanly *et al.*, "Multi-cell mimo cooperative networks: A new look at interference," *IEEE journal on selected areas in communications*, vol. 28, no. 9, pp. 1380–1408, 2010.

[156] X.-H. You, D.-M. Wang *et al.*, "Cooperative distributed antenna systems for mobile communications [coordinated and distributed mimo]," *IEEE Wireless Communications*, vol. 17, no. 3, pp. 35–43, 2010.

[157] F. Xia, Z. Wang, and Y. Sun, "Integrated computation, communication and control: Towards next revolution in information technology," in *International Conference on Intelligent Information Technology*. Springer, 2004, pp. 117–

125.

[158] J. Wang, C. Jiang *et al.*, "Thirty years of machine learning: The road to pareto-optimal wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1472–1514, 2020.

[159] L. Huang, S. Bi, and Y.-J. A. Zhang, "Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks," *IEEE Transactions on Mobile Computing*, vol. 19, no. 11, pp. 2581–2593, 2019.

Biographies