J Inf Process Syst, Vol.16, No.4, pp.896~914, August 2020
https://doi.org/10.3745/JIPS.03.0143

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Resource Management in 5G Mobile Networks: Survey and Challenges

Wei-Che Chien*, Shih-Yun Huang**, Chin-Feng Lai***, and Han-Chieh Chao**

Abstract

With the rapid growth of network traffic, a large number of connected devices, and higher application services, the traditional network is facing several challenges. In addition to improving the current network architecture and hardware specifications, effective resource management means the development trend of 5G. Although many existing potential technologies have been proposed to solve the some of 5G challenges, such as multiple-input multiple-output (MIMO), software-defined networking (SDN), network functions virtualization (NFV), edge computing, millimeter-wave, etc., research studies in 5G continue to enrich its function and move toward B5G mobile networks. In this paper, focusing on the resource allocation issues of 5G core networks and radio access networks, we address the latest technological developments and discuss the current challenges for resource management in 5G.

Keywords

Cloud Computing, Edge Computing, Network Slicing, Resource Management, 5G, 5G RAN Techniques

1. Introduction

The rapid development of the Internet of Things (IoT) [1] and high-quality multimedia [2] has led to an exponential increase in data traffic year by year. According to this trend, the cellular network will become more and more congested and the quality of service (QoS) of mobile users will be degraded. As traditional networks cannot carry these service requests, International Mobile Telecommunications-2020 (IMT2020) proposes three major directions for 5G application scenarios, including enhanced mobile broadband (eMBB), massive machine type communications (mMTC) and ultra-reliable low-latency communications (URLLC). The 3rd Generation Partnership Project (3GPP) develops 5G standard specifications based on these three directions too.

Many technologies and architectures also support the development of 5G. In order to reduce the increasing operating pressure of core network equipment, the European Telecommunications Standards Institute (ETSI) proposed the concept of edge computing [3-6]. The concept of edge computing and fog computing is to put some resources near the UE, such as the base station (BS). Therefore, the user can obtain the data through the BS and does not need to wait for a response from the remote server. In this

Manuscript received January 16, 2020; first revision April 23, 2020; accepted April 28, 2020.
Corresponding Author: Han-Chieh Chao (hcc@niu.edu.tw)
* Dept. of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan (b9944006@gmail.com)
** Dept. of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan (deantt67@gmail.com, hcc@niu.edu.tw)
*** Dept. of Engineering Science, National Cheng Kung University, Tainan, Taiwan (cinfon@ieee.org)

way, it can reduce transmission latency and obtain better QoS. The remote server does not have to service devices directly. It only needs to process a small amount of information sent by the BS, so it can reduce the burden on the core network. In addition, information-centric networking (ICN) [7-10] has some similar concepts to edge computing and fog computing. It caches data at neighboring network nodes to reduce the latency caused by the UE obtaining data from a remote server.

In order to increase the diverse application requirements of 5G, network slicing [11-14] is a must-have technology today. As early as February 2015, the "5G White Paper" proposed by the Next Generation Mobile Networks (NGMN) mentioned that the concept of network slicing technology was included in it [15]. In the 3GPP TR 22.891 technical report [16], the network slice is described as a collection of logical network functions that can support communication service requirements for specific application scenarios. At the ITU FG-IMT-2020 meeting, the participants also made comments on the network slicing technology, which also discussed the development of SDN and NFV in the 5G network architecture and the importance of network slicing technology.

Therefore, the combination of software-defined networking (SDN) [17-19] and network function virtualization (NFV) [20] is one of the current 5G development trends. SDN separates the control plane and data plane in the traditional network and sent control massage to many network devices through the controller. In this way, the network device only needs to process the packets, which greatly improves the control and the efficiency of network resources. NFV virtualizes the network functions of physical devices, such as firewalls, routers, deep packet inspection, and load balancers.

The implementation by software has excellent flexible configuration characteristics. In addition to increasing the efficiency of network service deployment, it can also reduce the purchase costs of hardware. SDN and NFV are independent of each other and can exist independently in the network system, but the two frameworks are highly complementary. For example, under the SDN architecture, when different services are provided to different users at the same time, customized services need to be deployed quickly. Therefore, NVF technology is required to achieve this; the establishment and communication between NFV require SDN support to achieve. Therefore, SDN restructures network architecture and does not change the functions of the network. NFV changes the type of network equipment without changing the functions of the equipment.

In addition to the 5G RAN system, the current mobile networks also coexist with different communication protocols such as 4G (LTE), WLAN (Wi-Fi), UMTS, and LoRa, gradually becoming a true World Wide Wireless Web (WWWW) system that is supported by millimeter-wave, filter bank multi-carrier, massive multiple-input multiple-output (MIMO) [21], D2D communication [22], co-frequency co-time full-duplex (CCFD) [23], ultra-dense networks (UDN) [24], and network slicing. In addition, 5G base stations are composed of macro cells and small cells (micro-BS, pico-BS, femto-BS). Some base stations use the cloud radio access network (C-RAN) architecture [25,26], which divides traditional BS into centralized unit (CU) and distribute unit (DU) to improve the flexibility of resource allocation. The core network has evolved into service-based architecture (SBA). On the whole, 5G will form a complex and flexible heterogeneous network architecture (HetNet) [27]. Fig. 1 shows a 5G architecture diagram.

With such huge network traffic and computational resource requirements, resource management (RM) is very important for 5G. RM can not only effectively improve the utilization of spectrum resources but also reduce transmission latency and save energy consumption through effective resource allocation. Therefore, the goal of this paper is to provide an introductory guide to the development of resource

allocation in 5G. We review RM in 5G of the existing works and provide a comprehensive classification of them. According to different network architecture, RM issues in 5G can be divided into two main categories: radio access network (RAN) and 5G core network (5GC). The goal of RM in RAN is mainly to improve spectrum efficiency (SE) and energy efficiency (EE). According to the communication type, the SE is divided into two sub-categories: cognitive radio network (CRN) and 5G. Among them, 5G RAN can be divided into fog radio access network (F-RAN) and C-RAN/H-RAN (hierarchical radio access network) according to the network architecture.

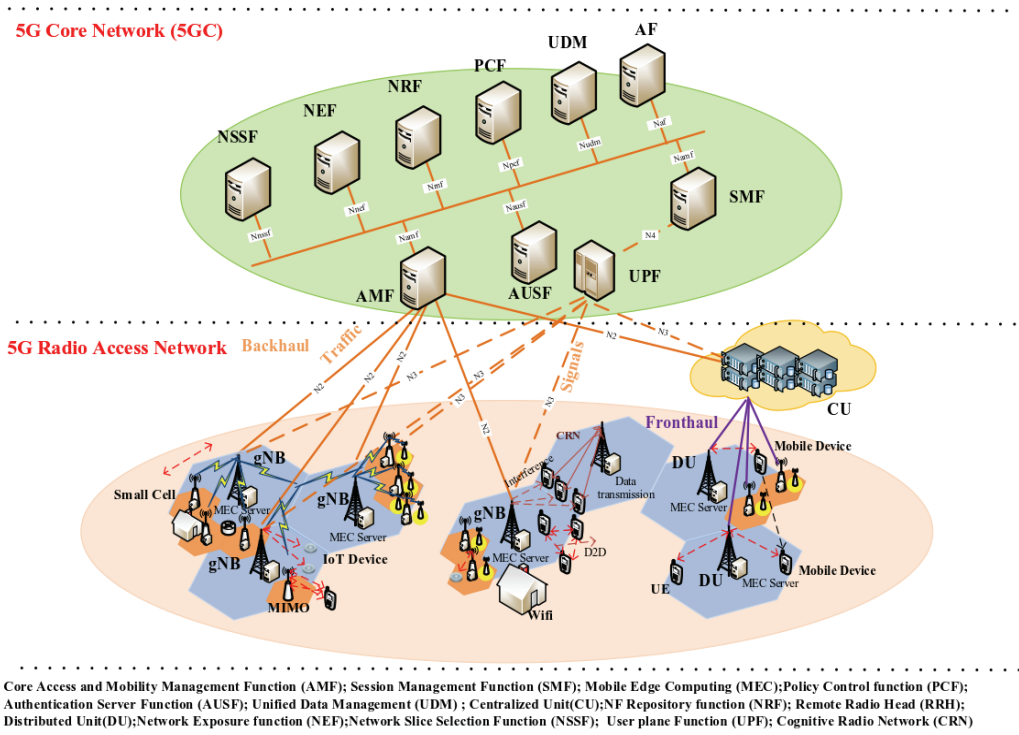


Fig. 1. The 5G architecture diagram.

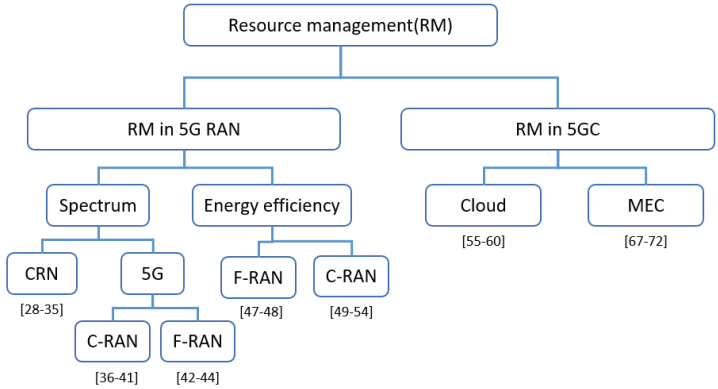


Fig. 2. Hierarchical classification of the resource allocation in 5G literature.

Besides, according to the core network architecture, RM in 5GC is divided into two sub-categories: RM in 5GC with cloud and RM in 5GC with mobile edge computing (MEC). The picture of the proposed hierarchical or graded classification is shown in Fig. 2 together with the associated literature. The main contributions can be summarized as follows:

- We review the recent development of resource allocation in 5G both radio access network and core network.
- We classify existing works from a multidimensional perspective, including application goals, service type, resource type, etc.
- We compare and analyze recent works and discuss their advantages and drawbacks.
- We discuss some open issues and challenges in 5G resource allocation, and outline some important future research directions.

The rest of the paper is organized as follows: Sections 2 and 3 introduces the development and the challenge of resource allocation in 5G RAN and core networks. In Section 4, we provide some important future research directions. Section 5 concludes this survey.

2. Resource Management in 5G Radio Access Network

At 5G RM for RAN, it can divide two parts to discuss. First is the spectrum resource, the second part is energy-efficient. There is some introduction to the literature as follows.

2.1 Resource Management with Spectrum

2.1.1 The main challenges

With the development of new modulation techniques and MIMO in 5G, radio resource allocation becomes more important. Although the CRN can help the UEs to dynamically access the licensed channel to enhance the spectrum efficiency, the interference for the primary user (PU) and secondary user (SU) cannot be ignored. On the other hand, the fairness problem still exists for spectrum sharing.

C-RAN and F-RAN are regarded as promising architectures for 5G. However, the fronthaul bandwidth in C-RAN is still the bottleneck for data transmission. Therefore, how to allocate resources between baseband unit (BBU) and remote radio head (RRH) is one of the challenges in C-RAN.

2.1.2 The conventional solutions

Since spectrum resources are limited, how to effectively manage wireless network resources has been widely discussed. Although 5G uses high frequencies to expand the use of bandwidth, coverage and interference are potential challenges. Recently, most research and scheme are taken into account the interference, transmission rate, throughput, and spectrum prices to achieve spectrum resource management (SRM). In this subsection, we introduce this research with different communication types, involving CRN, 5G (C-RAN and F-RAN), etc. The detection schemes for resource management with spectrum is captured in Table 1.

Cognitive radio network: The CRN is a promising solution for SRM. Spectrum can be divided into licensed bands and unlicensed bands. The licensed band only applies PUs to use. On the other hand,

unlicensed bands are not complete and hard to use. Therefore, the SU can sense and try to access the licensed band when SU find the idle channel slot. In this way, the spectrum resource can be used efficiently. The CRN is an important technique in 5G but it still needs to overcome some problems to improve network performance.

Table 1. The detection schemes for resource management with spectrum

Study	Year	Network architecture	Scheme	Important metrics	Simulator
[28]	2017	5G CRN	MIMO-OFDMA-based relaying	SINR, transmit power and transmission rate	MATLAB
[29]	2018	Enhanced 5G CRN	Dynamic spectrum aggregation	SINR, total shared licensed spectrum	N/A
[30]	2019	Spectrum access system	Need based spectrum allocation (NESA)	Bandgroup blocking probability	NS-2
[31]	2018	CRN	CCI mitigation SUs clustering algorithm and max-min utility optimization	Correct reception probability	N/A
[32]	2018	CRN	Load Estimator for Multiple Secondary Users	Number of selected channels	MATLAB
[33]	2018	Dynamic spectrum access network	Channel pricing algorithm	User preference, spectrum pricing	N/A
[34]	2018	CRN	Optimal monotonic optimization algorithm and suboptimal monotonic optimization algorithm	Average total throughput	N/A
[35]	2018	CRN	Deep neural network	Average spectral efficiency	N/A
[36]	2019	C-RAN	Spectrum efficiency-based joint optimization for offloading and resource allocation (SJOORA) scheme	Profit	N/A
[37]	2017	C-RAN	Interference-aware greedy heuristic algorithm	Sum rate	N/A
[38]	2019	C-RAN	Fixed point algorithm for channel allocation	Revenue	N/A
[39]	2019	5G RAN	Intelligent resource scheduling strategy (iRSS) for 5G RAN slicing	Revenue	Python
[40]	2018	C-RAN	Distributed mode selection and resource allocation+ CSI based many-to-many matching algorithm for RRH association+ Stackelberg game based D2D power control	Transmit power	N/A
[40]	2018	C-RAN	Dinkelbach-based algorithm	SINR fidelity	MATLAB
[42]	2019	FeRANs	Delay-aware bandwidth allocation	Migration traffic, non-migration traffic, and granted time slot	MATLAB
[43]	2019	F-RAN	Cluster formation algorithm, single-agent RL based caching, multi-agent RL based caching	System throughput, channel state information (CSI)	N/A
[44]	2019	F-RAN	GM-SPAS, MSRT-SPAS	Spectrum price	N/A
[45]	2020	IoT	Static scheduler and dynamic borrowing scheduler description	Number of resource blocks	MATLAB
[46]	2019	V2X network	IHG-RA	Received signal, SINR, RB	N/A

FeRANs=fog-enhanced radio access networks, GM-SPAS, game model based spectrum pricing and allocation scheme, MRST=multiple spectrum reuse technology based spectrum pricing and allocation scheme.

To increase the network capacity, the MIMO-OFDMA (orthogonal frequency-division multiple access) has been proposed in many studies. In [28], to guarantee the QoS with delay bound constriction, the authors have formulated the non-convex problem to convex problem for maximum effective capacity. To satisfy the objective function, they also proposed the scheme of heterogeneous statistical QoS-driven power allocation with mathematical. Although proposed method can efficiently reduce the power consumption and support high QoS, they only consider the SINR to define the problem.

In [29], the authors focus on enhanced 5G CRN (E-CRN). It means that these systems joint the spectrum sensing and geolocation by TV white space (TVWS). There is a trade-off problem, however, between spectrum sharing with licensed and spectrum aggregation with unlicensed. To solve the problem, the functions of shared value and the aggregate value are defined in this literature. In order to improve spectrum efficiency, this paper proposes dynamic spectrum aggregation and spectrum lean management.

At the same time, Xin et al. [30] also take into account the spectrum sharing. The difference from [29] is that the authors consider the spectrum sharing of PU and SU on geolocation. In order to solve the problem of frequency band resource allocation, the authors separately propose band allocation or release at the root spectrum access system (SAS) and sub-band allocation of streams in local SAS. They analyze the performance through mathematical proof. Although these papers almost consider the spectrum sharing with PU and SU and the network performance, they only take into account the interference in their system. Unfortunately, the latency is also an important metric in CRN. At the same time, these researches do not consider the fairness between PUs and SUs.

In CRN, as SU has the opportunity to access the licensed frequency band, spectrum resources can be effectively utilized. However, if there are multi-SUs in CRN, how to manage resources is important. Xu et al. [31] propose the fairness power allocation strategy. First, the authors consider the problem of maximum correct reception probability (CRP) is a non-convex and complex problem. Therefore, the authors divide the problem into two parts, namely sub-channel allocation and max-min utility power allocation. At the part of co-channel interference (CCI) sub-channel allocation, the main idea is following the average SINR between two SU, and the k-means will use for clustering the SUs into multiple groups. Next, the max-min utility power allocation the nonlinear Perron-Frobenius theory is considered for this problem. The authors take into account the interference between SUs and assume that the channel state information is perfect.

This assumption, however, does not fit the real environment situation. In [32], the authors proposed a load estimator for multiple SUs to find the largest useable spectrum hole. In this paper, the authors consider the length of the spectrum hole to find the maximum transmission rate. This estimating method can improve the transmission rate. Li et al. [33] consider the different qualities of the main system and formulate a spectrum pricing model based on the Hotelling game. SUs can buy channel according to the rule of Hotelling game. In this research, the authors also consider the interference of SUs with the primary system. To achieve the Nash equilibrium, an iterative pricing algorithm is proposed to solve this problem. Although this paper can find the sub-optimal solution, there is a high time complexity for this method.

The power constraints of SUs are considered in [34]. If the power of SUs is too large, it will cause the PU to suffer very serious interference, which is not fair to the PU. The problem of maximum SUs throughput and optimal sensing threshold, however, should be overcome in multi-band-multi-user. Hence, the monotonic based scheme was proposed in this paper. By using this approximate method can find the sub-optimal solution; nevertheless, find the global optimal also a problem. Hence, in [35], a deep neural network (DNN) is proposed to solve the spectrum resource allocation in multi-channel CRN. The

concept is to determine the ratio between total transmit power and allocated transmit power at each channel by allocating the transmit power of the SU. It is worth noting that DNN must consider the tradeoff problem between SUs transmits power and interference of PUs. Although deep learning can find better solutions than approximate methods, training time is an important consideration.

Cloud radio access network based: Since 2009, China Mobile has proposed the concept of C-RAN. The C-RAN is still the topical issue in the last ten years. As the characteristics of C-RAN centralized management, resource management can be very easy. Although CRN can effectively use spectrum resources, most operators do not support the use of CRN due to operators' profit factors. On the other hand, the fairness of PUs is a big problem. Therefore, some literature focuses on the C-RAN and hopes it can help allocate spectrum resources.

In [36], the task-aware C-RAN with MEC structure was considered. This structure can help the operators get the extract profit and process the task with high computation needs. The authors consider some metrics, including limited bandwidth at fronthaul, latency with offloading, and computational resources. They formulate SE based joint optimization for offloading and resource allocation problem as maximum profit, which is an NP-hard problem. The results show that both the number of successful offloading and the probability of successful offloading are almost close to the optimal solution. The proposed method, however, requires a long computational time. In [37], the authors formulate the problem of joint users to BS association and try to maximize the resource utilization of the network. To solve this problem, the authors propose the heuristic algorithm based on greedy for coordinated scheduling. In this paper, they reduce the computing complexity but when the number of users becomes more and more, the error with the sum rate becomes large.

At the same time, the concept of resource allocation for pricing has proposed in [38]. The mobile network operator (MNO) has the physical resource, and the mobile virtual network operator (MVNO) will lease the resource from MNO. Using lease resources from MNO, the MVNO can service the user. This problem, however, can be divided into two parts to discuss, namely maximizing the revenue from MVNO to MNO and maximizing the utility-surplus from MVNO to the user. Although the fixed-point algorithm can reduce the time complexity, the MNO maybe want to find the best revenue that spectrum utility becomes low.

Consider network slicing, a collaborative learning framework has proposed in [39]. The learning framework includes deep learning (DL) and reinforcement learning (RL). The main goal in this paper is to minimize the mean-square-error (MSE) and design schedule on-line resource.

Although the C-RAN can easily manage the resource, the device-to-device (D2D) communication may be one of the solutions to increase the SE and reduce the latency. Considering the loading of the BBU pool and fronthaul, the idea of adding D2D communication to the C-RAN has proposed in [40]. To find the maximum SE, the RL based strategy has used in this paper. Even though this proposed method can improve the system SE obviously, they only consider the uplink. In [41], the authors give a definition of the channel status in the C-RAN and evaluate its channel status information. In this paper, they prove the better SINR fidelity by increasing the transmit power for a UE but there are still problems with multi-users.

Fog radio access network based: In 5G, the F-RAN is one of the techniques to achieve the goal of URLLC. The concept of F-RAN is to join the communication infrastructure into fog computing. In this way, the computing component can be closer to UEs. Although F-RAN is one of the promising archi-

tecture for 5G, there is some problem that needs to be solved.

Li et al. [42] take into account the limited resources of F-RAN. Hence, F-RAN needs to migrate service to improve performance. To smooth the service migration, the authors consider the delay metrics and propose the bandwidth slicing scheme to dynamic allocation the bandwidth for service migration and non-migration. It is worth noting that this study only focuses on single-wavelength bandwidth slicing. The multi-wavelength bandwidth in the real world, however, is an important issue. In [43], the Stackelberg game was to be used to find the optimal cache resource and radio resource. In this literature, the role of resource management and fog access point (FAP) is a leader and follower, respectively.

To achieve a stable state, the cluster formation algorithm was designed for FAP. At the next step, the single-agent reinforcement learning (SARL) algorithm and multi-agent reinforcement learning (MARL) algorithm was proposed to find the global and local optima cache allocation. Simultaneously, the Stackelberg game also proposes in [44]. Difference between [43], the concept of spectrum pricing was taken into the game and formulate. In this way, the waste of spectrum resources can be avoided. On the other hand, the literature also considers spectrum reuse to improve spectrum utilization. In addition, the authors also discuss the relationship between the real situation of base station dynamic coverage and UEs requirements. These two papers also using game theory to improve system performance and achieve a win-win.

Other: Due to the mMTC requirement, IoT communication is an important issue in 5G. In [45], the authors focus on radio resource management (RRM) for multi-traffic IoT communication in 5G. In this paper, the authors discuss the human-to-human (H-H) and machine-to-machine (M-M) communication. And consider the maximum bandwidth utilization rate in one transmission time intervals (TTI). To solve the problem of maximum bandwidth utilization rate, two new scheduling methods have proposed, called static schedule and dynamic borrowing scheduler description.

At the initial step, the resource blocks will assign in one TTI, and the traffic classifier distinguishes the traffic that is H-H or M-M. When traffic has a classifier, the required number of a resource block for M-M or H-H will compare with the assigned resource block at the initial step. When the required RBs exceeds the usable RBs, it only accepts a part of RBs. Finally, the bandwidth utilization rate will be calculated. Even though this paper divides the traffic into M-M flows and H-H flows, IoT communication will become more complicated in the future. Therefore, more types of streams need to be considered. In addition to IoT, vehicle-to-everything (V2X) is also an important scenario [46]. To improve the sum rate, the authors proposed the interference hypergraph-based resource allocation (IHG-RA) scheme in non-orthogonal multiple access (NOMA)-V2X communication. Consider the situation with the transmit power of the vehicle, even though the sum rate has reduced, the proposed method can still keep the sum rate within acceptable.

2.2 Resource Management with Energy Efficiency

2.2.1 The main challenges

Recently, there are many literature focus on 5G RAN. The topic of EE is an emergency issue, and it needs to solve first. According to network architecture and requirements of 5G, the different network types and the massive number of devices should contain. There are some challenges to the EE issue. (1) In C-RAN, due to the feature of centralized management, the architecture of RRH-BBU and fronthaul

transmission. These features that computing power is increasing, as well as, the management is more easily. Consider that, however, when overloading of fronthaul/BBU pool and computing complex task; it will increase the energy consumption. (2) In the wireless network, the QoS is an important metric to analyze network performance. If we want to reduce energy consumption, however, there is a high probability of sacrifice the QoS. Therefore, the tradeoff problem between QoS and EE has to consider resource management.

2.2.2 The conventional solutions

The EE will discuss in this section. With the mobile device increase very quickly, energy consumption becomes an important issue. For the telecom operator, the energy cost is the heavy loading and unhappy to see it. Hence, most literature focus on reducing energy consumption. The detection schemes for resource management with EE is captured in Table 2.

Table 2. The detection schemes for resource management with energy-efficient

Study	Year	Network architecture	Scheme	Important metrics	Simulator
[47]	2019	F-RAN	DRL-based communication mode selection and resource management	Power consumption	N/A
[48]	2018	F-RAN	Iterative resource allocation algorithm	Economical energy efficiency (E3)	N/A
[49]	2017	C-RAN	Hierarchical location-based	Handover, host utilization	N/A
[50]	2018	C-RAN	Sleeping strategy	Power consumption, delay	N/A
[51]	2018	C-RAN	Clustering+ scheduling	Downlink throughput, RTT, processing time	Open-air-interface (OAI)
[52]	2018	H-CRAN	Online learning algorithm	BER, capacity, EE	Software defined radio testbed
[53]	2019	H-CRAN	Arrival rate based average energy-efficient	EE, queue length, power consumption	N/A
[54]	2019	H-CRAN	Iterative algorithm	Throughput, Utility function	N/A

Fog radio access network based: In [47], the authors consider the network is dynamic and assumes that the user can choose to use D2D communication or C-RAN. Due to this reason, resource allocation becomes more difficult. To achieve the goal of saving energy, authors take into account models of calculation, cache, and energy consumption. They define the problem according to these models. Next, the deep RL-based algorithm has proposed to help to select the communication mode and optimize resource allocation. The authors use cache to do one of the metrics in this paper; nevertheless, they do not consider the power consumption that cache data migration. Simultaneously, Yan et al. [48] also concerned about the F-RAN using the Economical Energy Efficiency (E3) metric, including throughput, cache capacity, fronthaul transmission, and energy consumption. To solve this problem, the iterative algorithm, power allocation algorithm, and resource allocation algorithm has proposed in this paper. Most of the 5G applications, however, has low latency requirement but latency and computing time is the lack to discuss in this paper.

Cloud radio access network based: In [49], according to the UE mobile state and location, the authors have proposed a location clustering algorithm based on manage resources to reduce energy consumption and latency. However, this study does not restrict the constraint of fronthaul. In [50], the authors formulate the EARTH model that the power consumption model can fit the C-RAN. Then, the sleep strategy model was proposed in this paper, called the logarithmic barrier method. By using this method, power consumption can be significantly reduced. For the problem definitions of these two studies, one is based on the number of handovers and the other is based on the delay. These factors, however, will affect QoS and power consumption, so for C-RAN, these factors need to be considered and resolved together.

As mentioned earlier, when the BBU pool is overloading, the energy consumption will increase. Hence, the resource allocation for the BBU pool is also important. In [51], the authors formulated the computing energy consumption according to testbed results. Next, the two sub-problems were discussed and solved, including BBU energy-aware resource allocation and allocation of bandwidth and power. Finally, the authors develop a corresponding method to solve them. This work is to obtain real data through the testbed and formulate their problem. How to build a large-scale with the real environment by testbed, however, is still a problem to be overcome.

In [52], in order to improve the BS capacity, bit error rate, and EE, the authors propose the online learning resource allocation model in H-CRAN. The model overcomes the curse of dimensionality and reduces the convergence time. However, since the proposed method will allocate resources according to the number of UEs with high QoS requirements, it is unfair for UEs with low QoS requirements. Therefore, how to find a mechanism to solve this trade-off problem is very important. At the same time, Zhang et al. [53] also consider the EE in H-CRAN. First, the authors define the power consumption model based on H-CRAN architecture according to the transmit power constraints, the minimum average data rate, and average power. In the second step, the Lyapunov optimization method has used in this paper.

Then, the matrix for optimal power allocation can be obtained. Finally, the problem can solve by the arrival rate based on average energy-efficient (ARAE). For OFDMA uplink in H-CRAN, Amani et al. [54] want to solve allocate the problem of RRH and FAP. They formulate the problem as a non-convex curve problem, which is a kind of NP-hard problem. To solve this problem, the iterative-based scheme has used to approach the optimal solution. Although the proposed method can increase the throughput and reduce energy consumption, the tidal effect of traffic should be considered when designing the offloading strategy.

3. Resource Management in 5G Core Networks

RM in 5G core networks can be divided into parts. The first part is RM with MEC. The second part is RM with cloud. The details are as follows:

3.1 Resource Management with Mobile Edge Computing

3.1.1 The main challenges

The combination of SDN, NFV, and MEC is increasingly being witnessed in the literature for achieving network scalability and deployment flexibility. But the complexity of the placement of virtual network functions (VNF) due to latency, computational resources, data rate capacity, and data rate will be a

challenge for resource management. For example, although deploy massive NFV in MEC can reduce transmission latency for service applications, it will increase the computational cost for MEC. Therefore, an effective method is needed to solve the trade-off problem.

3.1.2 The conventional solutions

With the popularity of various smart devices and IoT sensors, mobile network traffic is rapidly increasing. In order to increase network scalability and deployment flexibility, NFV and SDN have become the key technologies that enable the 5G network service. In addition, MEC can greatly reduce the transmission time for the UE to obtain resources from the cloud server. The combination of NFV, SDN and edge computing can meet the extreme requirements of new and diverse applications. The comparison of literature on resource management with MEC is captured in Table 3.

For MEC resource management based on NFC architecture, Shi et al. [55] proposed a method that combines the MDP and Bayesian learning approach to dynamically allocate cloud resources for NFV components. They use historical data to predict future resource reliability, which helps to enhance resource management performance. However, the proposed model isn't comprehensive because the authors didn't consider NFV component dependencies between multi-tenants. In [56], the authors proposed a fuzzy service offload decision mechanism (FSODM) algorithm to balance traffic load, which can respond faster to deployment microservices. They divided system architecture into upper (cloud), middle (resource monitoring and management) and lower layers (microservices). Microservices communicate with each other by RestAPI [57]; Controller Module uses Kubernetes' kubectl API [58]. The experimental results show that FSODM can auto-determine the number of microservices that need to be extended and maximize resource utilization. Nevertheless, the authors didn't consider the transmission delay.

For MEC resource management based on NFC and SDN architecture, Basta et al. [59] proposed three optimization models. They aim to minimize the network load cost and data center resources cost through finding the optimal placement of the data centers and the SDN and NFV mobile network functions. The contribution in this paper is that the authors consider the joint optimization of VNF function chains, SDN controllers and switches. The multi-objective model results in Pareto optimal solutions shows that the proposed model can achieve a balance between network load cost and data center resources cost. Song et al. [60] proposed a VNF-RACAG (VNF resource allocation scheme based on Context-Aware Grouping) technology.

They aim to minimize the end-to-end delay in edge networks and the number of transfers between clusters. The contribution of this paper is that the authors consider the location and the requirements of the user to deploy VNF within the edge network. Simulation results show that compared with WiNE [61] and PSwH [62] schemes, the proposed VNF-RACAG algorithm can obtain significant gain in end-to-end delay. Do and Kim [63] proposed a heuristic algorithm called UARG. This algorithm can build a usage track tree to efficiently distribute the standby functions over different deployment areas, such as a server, availability zone, center. They compared the proposed UARG to other greedy-based algorithms—bandwidth greedy (BWGR), availability greedy (AVGR), and zone greedy (ZOGR). Simulation results show that UARG has better performance than other baseline solutions in terms of computing and bandwidth resources. Ma et al. [64] proposed a management architecture for 5G core network SBA and a workload allocation algorithm. They use MNIP to formulate problems and consider bandwidth cost, processing delay, and energy cost to reduce network operating costs.

Table 3. Comparison of literature on resource management with MEC

Study	Technology adoption	Simulator	Methodology	Metrics	Objective
[55]	NFV	Dell PowerEdge Servers R720 and R810 with RHEL 6.5 operating system, WorkflowSim	MDP	Time constraint and execution cost	Minimize cost for entire NFV workflow
[56]	NFV	XenCenter 7.0.0, Kubernetes	Fuzzy-based method	CPU, memory	Load balancing traffic
[59]	SDN+NFV	Java framework	MILP	SDN traffic, latency, CPU	Minimizing the network load cost and data center resources cost
[60]	SDN+NFV	MATLAB	MILP, context-aware grouping algorithm	CPU, link capacity in the RAN	Minimize end-to-end delay
[63]	SDN+NFV	Python, PyCarm	INLP, UARG	Resource capacity of data center, bandwidth, computing resource	Minimize total WAN bandwidth and computing resource consumption
[64]	SDN+NFV	MATLAB	MNIP, workload allocation algorithm	Energy cost, delay	Minimize delay, bandwidth, and energy cost of cloud and MEC.

3.2 Resource Management with Cloud

3.2.1 The main challenges

For all kinds of network services, if operators always use the same network equipment, it will cause a lot of waste of resources. Therefore, Network Slicing is considered to be a very important network architecture for 5G, which allows multiple logical networks to run on shared physical network infrastructure. Each logical network is isolated and can provide customized network resources, such as bandwidth, latency, capacity, and so on. In addition to network resources, each logical network also includes some computing and storage resources. The architecture of network slicing can use NFV or SFC to achieve some specific network functions but there are also some challenges. The concept of the cloud is to centralize resources. Integrating NFV and SDN technologies can provide the required resources according to different services, saving a lot of unnecessary resource waste.

Even so, there are bottlenecks in the implementation of resource management. For example, user mobility will affect the carrying capacity of the node, resulting in complexity in management. Therefore, how to dynamically deploy VNF according to changes in demand is a major challenge for resource management in 5G with cloud. In addition, frequent dispatch resources will also cause other costs.

3.2.2 The conventional solutions

In this section, resource management technologies in 5G with cloud are introduced and studied. In order to improve the efficiency of resource management, the methods of NFV combine cache, per-group slicing, dynamic optical resource allocation have been proposed in the existing methods. The detailed comparison of these methods is captured in Table 4.

Table 4. Comparison of literature on resource management with cloud

Study	Technology adoption	Simulator	Methodology	Metrics	Objective
[67]	NFV, per-group slicing	OpenEPC [72]	Service-slice mapping algorithm	Latency, mobility, throughput, UE density	Reducing operation cost
[68]	NFV	-	FCFS	Capacities	Optimal OPEX and the optimal ratio between virtualized entities and legacy equipment
[69]	NFV, cache	MATLAB	ILP, CRO-based Algorithm	Power consumption of CPU device, memory device, network interface card, and caching	Maximizing the utilization of physical caching resource
[70]	NFV	-	RRAS algorithm	Processing, memory, I/O module, Storage	Minimizing resource black holes
[71]	SDN, NFV	NFV-LTE-EPC [73]	BAAS	Throughput, arrival rate, Number of active data plane.	Maximizing data plane throughput
[72]	SDON, NFV	Fedora V11.0, OpenDaylight	OSLB	BBR, TLD, RUR	Maximizing the total network throughput

The symbol "-" means unmentioned.
ILP=integer linear programming, FCFS=first come first serve, FCFSFA=first come first serve first available, CRO=chemical reaction optimization; BAAS=bit rate aware auto scaling; RRAS=reference resource affinity score; OSLB=optical switch load balancing.

Most existing research on RM with cloud computing is based on NFV and NFV+SDN architectures. For NFV architectures, since network slicing may cause a loss of multiplexing gain available, resources allocated to each slice will become isolated and exclusive. To solve this problem, Shimojo et al. [65] propose a concept called "per-group slicing" and a service-slice mapping algorithm that can efficiently automate service grouping and slice creation/accommodation to improve multiplexing gain and resource-usage efficiency. To evaluate the performance of the proposed algorithm, they test the effect of loss of multiplexing gain at SGWs, PGWs, and the physical server.

Experimental evaluation shows that the proposed algorithm can reduce both the number of slices and function wastage through selecting appropriate parameters. Nonetheless, the parameters in this paper depend on the actual functions for each service. It needs to quantify the thresholds to balance the number of slices and the total amount of function wastage.

It is important for resource management to find a trade-off between legacy and virtualized entities to be able to provide services and meet the QoS. Therefore, Abaev and Tsarev [66] proposed a hysteretic approach based on the queuing model for triggering scale out/in process. In addition, they consider the hybrid 5G EPC core, including legacy and virtualized entities. The simulation results show that the number of legacy entities can significantly increase system performance. In [67], the authors proposed an efficient caching resource allocation scheme in 5GC to maximize the utilization of physical caching resources. The contribution of this paper is that they propose a concept that integrates network slicing and in-network caching. ILP model is also used to formulate cache resource allocation problems.

The proposed scheme based on CRO [68] which mimicked the molecular interactions in the chemical reaction. Simulation results show that the concept of integrating cache and network slicing can improve the utilization of physical caching and save the cost of CapEx and OpEx. In [69], the authors proposed an approach that enables cloud infrastructure management and orchestration (MANO). They proposed a Reference Resource Affinity Score (RRAS) as a metric to optimize decisions and actions of the

virtualized systems inside the cloud infrastructure. But these methods didn't consider realistic traffic and system conditions.

For NFV+SDN architectures, Buyakar et al. [70] proposed an auto-scaling approach called bit rate aware auto scaling (BAAS) to solve problems that inappropriate network slicing may cause either over-provisioning or under-utilization of the underlying network infrastructure resources. The contributions in this paper include that authors implement a network slicing based 5G network architecture by extending NFV-LTE-EPC framework as a testbed and maintain the bit rates of data planes and number of data planes required. However, the proposed method can be improved by creating network slicing architecture with various verticals

In [71], the authors proposed a software defined optical networking (SDON) architecture and an optical switch load balancing (OSLB) algorithm for resource allocation. This architecture enables dynamic resource allocation according to the users' requirements. The simulation results show that the proposed algorithm is adaptable to the environment of mobile networks of handover and overload.

4. Open Issues and Future Trends

In the part of RM in 5G RAN, there are some issues that need to be solved. On the spectrum, dynamic spectrum allocation is a popular issue. At this issue, the method of spectrum sensing and spectrum sharing in CRN is very important. When PUs leaves the licensed channel to find the idle time slot, and the SUs can access licensed channel to achieve the spectrum sharing. Although the spectrum sharing and dynamic spectrum access can achieve the maximum SE, the SUs will make the interference to PUs also need to consider. On the other hand, the MIMO and new radio modulation technology is an important issue in 5G RAN. Hence, most of the literature used network slicing to help solve this problem.

In addition to spectrum resource allocation, the EE is also important in 5G. According to the network architecture for 5G, it is complex than the traditional cellular network. Due to this reason, it will also affect the power consumption increase. At the EE, the BBU pool resource allocation is a problem in C-RAN. Most of the paper has been formulated the problem can meet the game theory and using the iterative algorithm to find the sub-optimal solution. On the other hand, F-RAN also is a good idea to solve the disadvantage of centralized management. However, the F-RAN also will be faced with the problem, such as caching capacity, throughput. Therefore, how to offload or select the cache node is an important issue in F-RAN. At the same time, joint the D2D communication in C-RAN or F-RAN may be a solution to improve the SE. But using D2D communication has a high probability increase in energy consumption. Hence, how to allocate the bandwidth resource and power is one of the issues in 5G RAN.

In the part of RM in 5GC, the combination of network slicing and MEC is the current network architecture trend. Operators can deploy some VNFs into edge servers to reduce the delay of obtaining data from the remote cloud server. Compared with cloud architecture, MEC can provide more immediate services. However, how to achieve energy-saving and high utilization efficiency deployment is a challenge. In addition, as user mobility affects bearer and resource allocation of VNF, dynamic and immediate resource allocation needs to be considered to maximize 5G resource efficiency.

We review the resource allocation in 5G RAN and 5GC and provide some trends we observed. First, in order to satisfy the three requirements in 5G RAN, the relationship between the licensed band and the unlicensed band should be considered. By doing this way, we can utilize the spectrum more efficiently. Second, the H-CRAN or F-RAN will be a promising network architecture in 5G. On the other hand, the

D2D communication may be joint into the H-CRAN or F-RAN to enhance the SE. Third, network slicing is an important technique that dynamic spectrum and computing resource allocation in 5G RAN and 5GC can be implemented. Fourth, there are more and more research using learning algorithm, such as RL, long short-term memory (LSTM), and convolutional neural network (CNN), to make resource allocation more intelligence. Fifth, integrating edge computing and cloud computing to reduce computational time for real-time resource allocation is a trend for complex network architecture.

5. Conclusions

Multimedia, augmented reality (AR), virtual reality (VR), IoT, and other applications are continually developing. Traditional networks cannot support current network requirements. It is also difficult to solve the bandwidth and data requirements of a large number of UEs by upgrading the hardware specifications. Therefore, effective resource management is a promising solution for limited resources, including frequency resources, computing resources, and deployment costs. To achieve high elasticity and high flexibility resource management, network slicing is an indispensable technology. In addition, MEC has become one of the necessary network architecture to achieve complex and difficult application scenarios. This study investigates the resources management in 5G, including the core network and RAN, and systematically follows the network architecture, application scenarios, and research goals to classify recent works. According to the survey, challenges for related research is proposed. We also provide readers with some future research trends and hope to promote more researchers to master 5G resource allocation.

Acknowledgement

This paper was supported in part by the Ministry of Science and Technology of Taiwan (No. MOST 107-2221-E259-005-MY3).

References

[1] S. C. Wang, W. S. Hsiung, K. Q. Yan, and Y. T. Tsai, "Optimal agreement achievement in a fog computing based IoT," *Journal of Internet Technology*, vol. 20, no. 6, pp. 1767-1779, 2019.

[2] C. Zhang, H. H. Cho, C. Y. Chen, T. K. Shih, and H. C. Chao, "Fuzzy-based 3-D stream traffic lightweighting over mobile P2P network," *IEEE Systems Journal*, vol. 14, no. 2, pp. 1840-1851, 2020.

[3] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637-646, 2016.

[4] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78-81, 2016.

[5] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, *Mobile Edge Computing: A Key Technology Towards 5G*. Sophia Antipolis, France: European Telecommunications Standards Institute, 2015.

[6] T. H. Luan, L. Gao, Z. Li, Y. Xiang, G. Wei, and L. Sun, "Fog computing: focusing on mobile users at the edge," 2016 [Online]. Available: https://arxiv.org/abs/1502.01815.

[7] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 26-36, 2012.

[8] I. Psaras, W. K. Chai, and G. Pavlou, "Probabilistic in-network caching for information-centric networks," in *Proceedings of the second edition of the ICN workshop on Information-centric networking*, Helsinki, Finland, 2012, pp. 55-60.

[9] G. Xylomenos, C. N. Ververidis, V. A. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. V. Katsaros, and G. C. Polyzos, "A survey of information-centric networking research," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 1024-1049, 2013.

[10] H. C. Chao, W. J. Jian, H. H. Cho, C. W. Tsai, and J. S. Pan, "Prediction-based cache adaptation for named data networking," *Journal of Computers*, vol. 27, no, 1, pp. 45-55, 2016.

[11] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94-100, 2017.

[12] NGMN Alliance, "Description of network slicing concept," 2016 [Online]. Available: https://www.ngmn.org/wp-content/uploads/160113_NGMN_Network_Slicing_v1_0.pdf.

[13] H. Zhang, N. Liu, X. Chu, K. Long, A. H. Aghvami, and V. C. M. Leung, "Network slicing based 5G and future mobile networks: mobility, resource management, and challenges," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138-145, 2017.

[14] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80-87, 2017.

[15] Next Generation Mobile Networks, "5G White Paper," 2015 [Online]. Available: http://ngmn.org/5g-white-paper/5g-white-paper.html.

[16] 3rd Generation Partnership Project (3GPP), "Feasibility study on new services and markets technology enablers," 3GPP Organizational Partners, *Technical Report TR 22.891*, 2015.

[17] A. C. Baktir, A. Ozgovde, and C. Ersoy, "How can edge computing benefit from software-defined networking: a survey, use cases, and future directions," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2359-2391, 2017.

[18] W. C. Chien, C. F. Lai, H. H. Cho, and H. C. Chao, "A SDN-SFC-based service-oriented load balancing for the IoT applications," *Journal of Network and Computer Applications*, vol. 114, pp. 88-97, 2018.

[19] W. C. Chien, H. Y. Weng, C. F. Lai, Z. Fan, H. C. Chao, and Y. Hu, "A SFC-based access point switching mechanism for Software-Defined Wireless Network in IoV," *Future Generation Computer Systems*, vol. 98, pp. 577-585, 2019.

[20] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, "Network function virtualization in 5G," *IEEE Communications Magazine*, vol. 54, no. 4, pp. 84-91, 2016.

[21] E. Bjornson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 114-123, 2016.

[22] P. Gandotra, R. Kumar Jha, and S. Jain, "A survey on device-to-device (D2D) communication: architecture and security issues," *Journal of Network and Computer Applications*, vol. 78, pp. 9-29, 2017.

[23] J. Li, H. Zhang, and M. Fan, "Digital self-interference cancellation based on independent component analysis for co-time co-frequency full-duplex communication systems," *IEEE Access*, vol. 5, pp. 10222-10231, 2017.

[24] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2522-2545, 2016.

[25] W. C. Chien, C. F. Lai, and H. C. Chao, "Dynamic resource prediction and allocation in C-RAN with edge artificial intelligence," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4306-4314, 2019.

[26] I. A. Alimi, A. L. Teixeira, and P. P. Monteiro, "Toward an efficient C-RAN optical fronthaul for the future networks: a tutorial on technologies, requirements, challenges, and solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 708-769, 2018.

[27] H. Ramazanali, A. Mesodiakaki, A. Vinel, and C. Verikoukis, "Survey of user association in 5G HetNets," in *Proceedings of 2016 8th IEEE Latin-American Conference on Communications (LATINCOM)*, Medellin, Colombia, 2016, pp. 1-6.

[28] X. Zhang and J. Wang, "Heterogeneous QoS-driven resource allocation over MIMO-OFDMA based 5G cognitive radio networks," in *Proceedings of 2017 IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, CA, 2017, pp. 1-6.

[29] W. Zhang, C. X. Wang, X. Ge, and Y. Chen, "Enhanced 5G cognitive radio networks based on spectrum sharing and spectrum aggregation," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6304-6316, 2018.

[30] C. Xin, P. Paul, M. Song, and Q. Gu, "On dynamic spectrum allocation in geo-location spectrum sharing systems," *IEEE Transactions on Mobile Computing*, vol. 18, no. 4, pp. 923-933, 2019.
[31] W. Xu, R. Qiu, and J. Cheng, "Fair optimal resource allocation in cognitive radio networks with co-channel interference mitigation," *IEEE Access*, vol. 6, pp. 37418-37429, 2018.
[32] S. Khodadadi, D. Qiu, and Y. R. Shayan, "Performance analysis of secondary users in cognitive radio networks with dynamic spectrum allocation," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1684-1687, 2018.
[33] F. Li, Z. Sheng, J. Hua, and L. Wang, "Preference-based spectrum pricing in dynamic spectrum access networks," *IEEE Transactions on Services Computing*, vol. 11, no. 6, pp. 922-935, 2018.
[34] X. Wang, S. Ekin, and E. Serpedin, "Joint spectrum sensing and resource allocation in multi-band-multi-user cognitive radio networks," *IEEE Transactions on Communications*, vol. 66, no. 8, pp. 3281-3293, 2018.
[35] W. Lee, "Resource allocation for multi-channel underlay cognitive radio network based on deep neural network," *IEEE Communications Letters*, vol. 22, no, 9, pp. 1942-1945, 2018.
[36] Z. Jian, W. Muqing, and Z. Min, "Joint computation offloading and resource allocation in C-RAN with MEC based on spectrum efficiency," *IEEE Access*, vol. 7, pp. 79056-79068, 2019.
[37] M. Awais, A. Ahmed, M. Naeem, M. Iqbal, W. Ejaz, A. Anpalagan, and H. S. Kim, "Efficient joint user association and resource allocation for cloud radio access networks," *IEEE Access*, vol. 5, pp. 1439-1448, 2017.
[38] J. Ye and Y. J. Zhang, "Pricing-based resource allocation in virtualized cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7096-7107, 2019.
[39] M. Yan, G. Feng, J. Zhou, Y. Sun, and Y. C. Liang, "Intelligent resource scheduling for 5G radio access network slicing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7691-7703, 2019.
[40] Y. Sun, M. Peng, and H. V. Poor, "A distributed approach to improving spectral efficiency in uplink device-to-device-enabled cloud radio access networks," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6511-6526, 2018.
[41] D. Chen, Z. Zhao, Z. Mao, and M. Peng, "Channel matrix sparsity with imperfect channel state information in cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1363-1374, 2018.
[42] J. Li, X. Shen, L. Chen, J. Ou, L. Wosinska, and J. Chen, "Delay-aware bandwidth slicing for service migration in mobile backhaul networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 11, no. 4, pp. B1-B9, 2019.
[43] Y. Sun, M. Peng, and S. Mao, "A game-theoretic approach to cache and radio resource management in fog radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 10145-10159, 2019.
[44] Y. Yu, S. Liu, Z. Tian, and S. Wang, "A dynamic distributed spectrum allocation mechanism based on game model in fog radio access networks," *China Communications*, vol. 16, no. 3, pp. 12-21, 2019.
[45] A. Saddoud, W. Doghri, E. Charfi, and L. C. Fourati, "5G radio resource management approach for multi-traffic IoT communications," *Computer Networks*, vol. 166, article no. 106936, 2020.
[46] C. Chen, B. Wang, and R. Zhang, "Interference hypergraph-based resource allocation (IHG-RA) for NOMA-integrated V2X networks," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 161-170, 2019.
[47] Y. Sun, M. Peng, and S. Mao, "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1960-1971, 2019.
[48] Z. Yan, M. Peng, and M. Daneshmand, "Cost-aware resource allocation for optimization of energy efficiency in fog radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2581-2590, 2018.
[49] U. Karneyenka, K. Mohta, and M. Moh, "Location and mobility aware resource management for 5G cloud radio access networks," in *Proceedings of the 2017 International Conference on High Performance Computing & Simulation (HPCS)*, Genoa, Italy, 2017, pp. 168-175.
[50] J. Luo, Q. Chen, and L. Tang, "Reducing power consumption by joint sleeping strategy and power control in delay-aware C-RAN," *IEEE Access*, vol. 6, pp. 14655-14667, 2018.
[51] A. Younis, T. X. Tran, and D. Pompili, "Bandwidth and energy-aware resource allocation for cloud radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6487-6500, 2018.

[52] I. Alqerm and B. Shihada, "Sophisticated online learning scheme for green resource allocation in 5G heterogeneous cloud radio access networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 10, pp. 2423-2437, 2018.

[53] Y. Zhang, G. Wu, L. Deng, and J. Fu, "Arrival rate-based average energy-efficient resource allocation for 5G heterogeneous cloud RAN," *IEEE Access*, vol. 7, pp. 136332-136342, 2019.

[54] N. Amani, H. Pedram, H. Taheri, and S. Parsaeefard, "Energy-efficient resource allocation in heterogeneous cloud radio access networks via BBU offloading," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1365-1377, 2019.

[55] R. Shi, J. Zhang, W. Chu, Q. Bao, X. Jin, C. Gong, Q. Zhu, C. Yu, and S. Rosenberg, "MDP and machine learning-based cost-optimization of dynamic resource allocation for network function virtualization," in *Proceedings of 2015 IEEE International Conference on Services Computing*, New York, NY, 2015, pp. 65-73.

[56] C. C. Liu, C. C. Huang, C. W. Tseng, Y. T. Yang, and L. Chou, "Service resource management in edge computing based on microservices," in *Proceedings of 2019 IEEE International Conference on Smart Internet of Things (SmartIoT)*, Tianjin, China, 2019, pp. 388-392.

[57] Kubernetes, "kube-proxy," 2020 [Online]. Available: https://kubernetes.io/docs/reference/command-line-tools-reference/kube-proxy.

[58] Kubernetes, "The Kubernetes API," 2020 [Online]. Available: https://kubernetes.io/docs/concepts/overview/kubernetes-api.

[59] A. Basta, A. Blenk, K. Hoffmann, H. J. Morper, M. Hoffmann, and W. Kellerer, "Towards a cost optimal design for a 5G mobile core network based on SDN and NFV," *IEEE Transactions on Network and Service Management*, vol. 14, no. 4, pp. 1061-1075, 2017.

[60] S. Song, C. Lee, H. Cho, G. Lim, and J. M. Chung, "Clustered virtualized network functions resource allocation based on context-aware grouping in 5G edge networks," *IEEE Transactions on Mobile Computing*, vol. 19, no. 5, pp. 1072-1083, 2020.

[61] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, "Scheduling wireless virtual networks functions," *IEEE Transactions on Network and Service Management*, vol. 13, no. 2, pp. 240-252, 2016.

[62] J. Plachy, Z. Becvar, and P. Mach, "Path selection enabling user mobility and efficient distribution of data for computation at the edge of mobile network," *Computer Networks*, vol. 108, pp. 357-370, 2016.

[63] T. X. Do and Y. Kim, "Usage-aware protection plan for state management functions in service-based 5G core network," *IEEE Access*, vol. 6, pp. 36906-36915, 2018.

[64] L. Ma, X. Wen, L. Wang, Z. Lu, and R. Knopp, "An SDN/NFV based framework for management and deployment of service based 5G core network," *China Communications*, vol. 15, no. 10, pp. 86-98, 2018.

[65] T. Shimojo, M. R. Sama, A. Khan, and S. Iwashina, "Cost-efficient method for managing network slices in a multi-service 5G core network," in *Proceedings of 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, Lisbon, Portugal, 2017, pp. 1121-1126.

[66] P. Abaev and A. Tsarev, "Hysteretic mechanism for 5G hybrid evolved packet core resource management," in *Proceedings of 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Moscow, Russia, 2018, pp. 1-6.

[67] Q. Jia, R. Xie, T. Huang, J. Liu, and Y. Liu, "Efficient caching resource allocation for network slicing in 5G core network," *IET Communications*, vol. 11, no. 18, pp. 2792-2799, 2017.

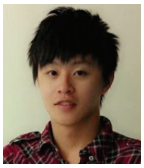
[68] A. Y. S. Lam and V. O. K. Li, "Chemical-reaction-inspired metaheuristic for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 14, no. 3, pp. 381-399, 2010.

[69] F. Z. Yousaf and T. Taleb, "Fine-grained resource-aware virtual network function management for 5G carrier cloud," *IEEE Network*, vol. 30, no. 2, pp. 110-115, 2016.

[70] T. V. K. Buyakar, A. K. Rangisetti, A. A. Franklin, and B. R. Tamma, "Auto scaling of data plane VNFs in 5G networks," in *Proceedings of the 2017 13th International Conference on Network and Service Management (CNSM),* Tokyo, Japan, 2017, pp. 1-4.

[71] Y. Zhao, Z. Chen, J. Zhang, and X. Wang, "Dynamic optical resource allocation for mobile core networks with software defined elastic optical networking," *Optics Express*, vol. 24, no. 15, pp. 16659-16673, 2016.

[72] The OpenEPC Project [Online]. Available: https://sites.google.com/a/corenetdynamics.com/openepc/project-info/open-source.
[73] NFV-LTE-EPC [Online]. https://github.com/networkedsystemsIITB/NFV_LTE_EPC.



Wei-Che Chien **https://orcid.org/0000-0002-0172-9958**

He is an assistant professor at Department of Computer Science and Information Engineering, National Dong Hwa University since 2020. He received his B.S. and M.S. degree in Computer Science and Information Engineering from National I-Lan University, Taiwan in 2014 and 2016, respectively. He is currently pursuing the PhD degree at the Engineering Science, National Cheng Kung University. His research interests include wireless rechargeable sensor networks, 5G mobile networks, IoT application, fog computing and cloud computing.

Shih-Yun Huang **https://orcid.org/0000-0002-6524-2110**

He received the B.S. and M.S. degree in Electrical and Electronic Engineering, National Ilan University, in 2009 and 2013, respectively. He is currently pursuing his PhD degree in electrical engineering at National Dong Hwa University. His research interests include Smart grid, mobile networks, cloud computing and mobile edge computing.

Chin-Feng Lai **https://orcid.org/0000-0001-7138-0272**

Hei is an associate professor at Department of Engineering Science, National Cheng Kung University since 2016. He received the Ph.D. degree in department of engineering science from the National Cheng Kung University, Taiwan, in 2008. He received Best Paper Award from IEEE 17th CCSE, 2014 International Conference on Cloud Computing, IEEE 10th EUC, IEEE 12th CIT. He has more than 100 paper publications. He is an associate editor-in-chief for *Journal of Internet Technology*. His research focuses on Internet of Things, body sensor networks, e-healthcare, mobile cloud computing, cloud-assisted multimedia network, embedded systems, etc. He is an IEEE Senior Member since 2014.

Han-Chieh Chao **https://orcid.org/0000-0003-2540-9200**

He received his M.S. and Ph.D. degrees in Electrical Engineering from Purdue University in 1989 and 1993, respectively. He is currently a Professor with the Department of Electrical Engineering, National Dong Hwa University, where he also serves as the President. He is also with the Department of Computer Science and Information Engineering and the Department of Electronic Engineering, National Ilan University, Taiwan. He is a fellow of IET (IEE) and a Chartered Fellow of the British Computer Society. He serves as the Editor-in-Chief for the *Institution of Engineering and Technology Networks*, the *Journal of Internet Technology*, the *International Journal of Internet Protocol Technology*, and the *International Journal of Ad Hoc and Ubiquitous Computing*.