Resource Allocation for Network Slicing in 5G Telecommunication Networks: A Survey of Principles and Models

Ruoyu Su, Dengyin Zhang, R. Venkatesan, Zijun Gong, Cheng Li, Fei Ding, Fan Jiang, and Ziyang Zhu

Abstract

With the rapid and sustained growth of network demands, 5G telecommunication networks are expected to provide flexible, scalable, and resilient communication and network services, not only for traditional network operators, but also for vertical industries, OTT, and third parties to satisfy their different requirements. Network slicing is a promising technology to establish customized end-to-end logic networks comprising dedicated and shared resources. By leveraging SDN and NFV, network slices associated with resources can be tailored to satisfy diverse QoS and SLA. Resource allocation of network slicing plays a pivotal role in load balancing, resource utilization, and networking performance. In this article, we focus on the principles and models of resource allocation algorithms in 5G network slicing. We first introduce the basic ideas of the SDN and NFV with their roles in network slicing. The MO architecture of network slicing is also studied, which provides a fundamental framework of resource allocation algorithms. Then, resource types with corresponding isolation levels in RAN slicing and CN slicing are analyzed, respectively. Furthermore, we categorize the mathematical models of resource allocation algorithms based on their objectives and elaborate them with typical examples. Finally, open research issues are identified with potential solutions.

Introduction

The emerging fifth generation telecommunication networks (5G) are conceived to offer a large amount of end-to-end network services for diverse requirements. These requirements are created by not only traditional mobile communication applications, but also vertical market segments, such as automatic driving, unmanned aerial vehicles, telemedicine, massive Internet of Things (mIoT), and so on [1]. To serve different application scenarios, verticals may require distinct network services by means of differentiated quality of service (QoS), service level agreements (SLAs), and key performance indicators (KPIs). In this case, the "one-size-fits-all" and "one-network-fits-all" types of fourth generation (4G) telecommunication networks (i.e., all network devices follow the same pipeline with few considerations on service customization) is no longer suitable, which are evolved into customized network paradigms by leveraging network slicing.

As one of the promising techniques in 5G, network slicing creates end-to-end (i.e., from radio access network (RAN) to core network (CN)) logically self-contained networks over the infrastructure of telecommunication networks [2]. Each logically self-contained network, referred to as a network slice, is designed for a specific requirement and consists of several network functions and resources abstracted from underlying communication and network resources. The fundamental idea of network slicing is closely related to that of infrastructure as a service (IaaS) in cloud computing. IaaS shares computing, storage, and networking resources among different tenants and provides fully-functional virtual networks powered by software defined networks (SDN) and network function virtualization (NFV) [3]. In fact, SDN and NFV are considered as the key technologies of network slicing in the context of 5G. SDN relies on the separation between the control plane and the data plane to enhances data forwarding efficiency and network programmability. NFV enables each network function, named virtual network function (VNF), to run on general-purpose hardware to alleviate deployment costs. In this case, a network slice, as a combination of several VNFs, not only provides a flexible, scalable, and programmable network service, but also reduces capital expenditures (CAPEX) and operational expenditures (OPEX) by efficiently orchestrating and managing VNFs [2].

As a main issue in 5G telecommunication networks, the resource allocation of network slicing faces various challenges in terms of isolation, customization, elasticity, and end-to-end coordination. Specifically, in the aspect of resource isolation, sharing and isolating resources for network slicing are not straightforward due to the varying communication environment. For instance, the stringent isolation of radio resources may result in low multiplexing gain. The customization enables the efficient resource utilization of network slicing to satisfy particular network service requirements. How to effectively convert the network service requirements to the desired network resources requires more consideration in different levels, including the control level, data plane level, and

Digital Object Identifier:
10.1109/MNET.2019.1900024

Ruoyu Su, Dengyin Zhang, and Fei Ding are with Nanjing University of Posts and Telecommunications; R. Venkatesan, Zijun Gong, Cheng Li, and Fan Jiang are with Memorial University of Newfoundland; Ziyang Zhu is with the University of Sheffield.

0890-8044/19/$25.00 © 2019 IEEE

network wide level. Moreover, the elasticity of network slicing is a challenging approach of adjusting resources under varying network conditions to guarantee the required SLA. For example, the exact number of virtual machines (VMs) with the corresponding computing, storage, and networking resources cannot be directly obtained under time-varying data traffic [2]. The dynamic allocation of shared resources may affect the network performance of network slices. End-to-end resource allocation has to cross different domains, such as CN, RAN, and transport network (TN). It is not easy to achieve a novel coordination among heterogenous techniques of different network layers.

Most surveys on network slicing investigate the diverse business and academic motivations in 5G, the fundamental architecture of a network slice, the principles of network slicing, the effect of SDN and NFV on next generation networks, and the algorithms of network slicing [4, 5]. More specifically, for RAN slicing, the specific requirements of RAN as well as the integration of fronthaul and backhaul networks are also discussed. The virtualization and the architecture of the next generation CN are always emphasized for CN slicing. Distinct from these surveys, our interests in this article focus on the principles and mathematical models of resource allocation algorithms for 5G network slicing. At first, we study the principles of SDN and NFV with their effects on network slicing. The fundamental management and orchestration (MO) architecture of network slicing is also presented to provide a comprehensive background knowledge of network slicing. Then, we analyze resource types and corresponding isolation levels in RAN and CN slicing respectively, which are the preliminaries of resource allocation algorithms. To the best of our knowledge, most resource allocation problems of network slicing can be converted to general optimization problems (we use a general model to represent the general optimization problem). For more complicated scenarios caused by diverse network environments and complex network slice requirements, several sophisticated and mature mathematical models are employed in the framework of the general optimization problem. Based on different mathematical approaches, we categorize these models into game theory based models, prediction techniques, and robustness/failure recovery models. We explain each mathematical model with typical examples. In addition, several open issues of network slicing are proposed with potential solutions. We believe that innovative resource allocation algorithms related to network slicing can be inspired by the analysis of these mathematical models.

The rest of this article is organized as follows. In the following section, we present the basic ideas of the SDN and NFV and the standard MO architecture of network slicing. Then we investigate the resource types and the isolation levels for both RAN and CN slicing. We elaborate the mathematical models of the resource allocation algorithms with examples following that. Future research directions and challenges are then discussed and we conclude this article in the final section.

The fundamental issues of network slicing include the description of a network slice requirement and the deployment of a network slice with its lifecycle management. To address these issues, the Third Generation Partnership Project proposes a fundamental MO architecture with corresponding network functions.

Background

Software Defined Networking and Network Function Virtualization

In the framework of SDN, the control plane is fully separated from the data plane, which is moved to a centralized location implemented by SDN controllers [3]. Based on the requirements of the application at hand, the SDN controllers generate different rules in terms of link discovery, topology management, policy deployment, and flow table delivery and send them to the data plane. The forwarding devices in the data plane, such as switches and routers, just apply and execute these rules. Unlike the tight coupling between network functions and proprietary hardware in conventional telecommunication devices, NFV enables network functions to run on general-purpose servers in the form of software appliances [3]. To efficiently and flexibly utilize virtual resources and manage VNFs in telecommunication networks, the NFV management and orchestrator (NFV-MANO) is proposed by the European Telecommunication Standards Institute (ETSI), which consists of NFV orchestration (NFVO), VNF managers (VNFMs), and virtualized infrastructure managers (VIMs). NFV-MANO manages the lifecycle of VNFs through VNFMs and VIMs. NFVO is responsible for orchestrating a network service incorporated with an external operation/business support system (BSS/OSS).[1]

SDN and NFV are considered as the key enablers for network slicing due to their programmability, scalability, and flexibility. For example, the evolved packet core (EPC) realizes the separation between the control plane and the user plane inspired by SDN, which leads to the flexible deployment of network functions. Moreover, network slices can be initialized and modified at low cost because NFV-MANO manages the creation, updating, and termination of network functions by VMs or containers. The NFVO can adjust resource allocation for VNFs incorporated with VNFM and VIM in varying network environments, such as fluctuation of data traffic, variation of users in network slices, and so on. In addition, when some VNFs fail, the routers and switches in the data plane of SDN can reroute data traffic based on the flow table or scheduling policy configured by SDN controllers.

Management and Orchestration Architecture of Network Slicing

The fundamental issues of network slicing include the description of a network slice requirement and the deployment of a network slice with its lifecycle management. To address these issues, the Third Generation Partnership Project (3GPP) proposes a fundamental MO architecture with corresponding network functions, including communication service management function (CSMF), network slice management function (NSMF), and network slice subnet management function (NSSMF) (i.e., each subnet has a NSSMF, such

[1] The NFVO allows a network operator to provide a network service by chaining a number of VNFs over physical resources

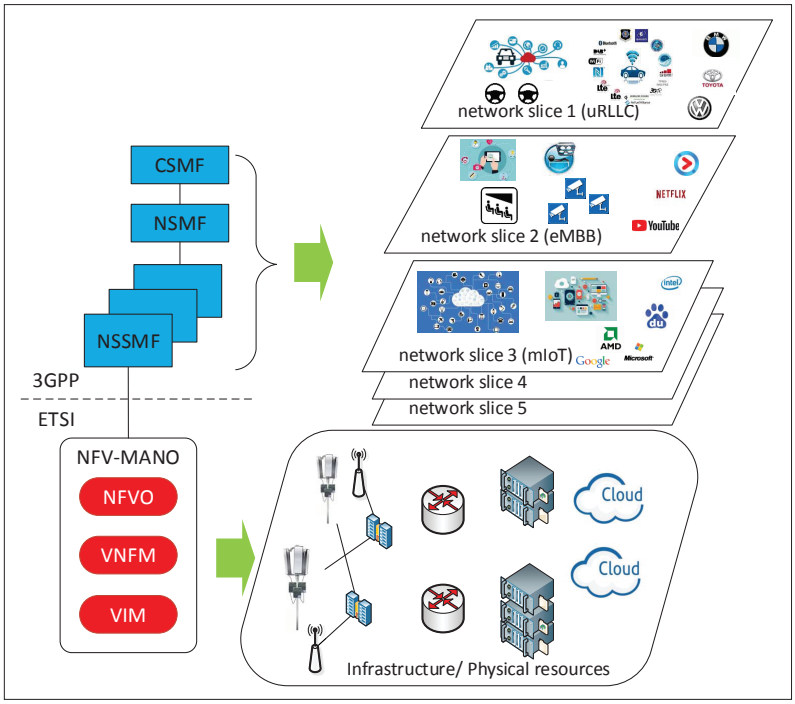

FIGURE 1. The MO architecture of network slicing.

as RAN, CN). The MO architecture of network slicing is presented in Fig. 1.

Under this architecture, the requirement of a specific network service is converted to the requirement of a network slice (i.e., network slice requirement description) by CMSF, which is delivered to NSMF. Then, the requirement of the network slice is decomposed into the requirements for different subnets by NSMF. NSSMFs in different subnets manage and orchestrate the network slice in the corresponding subnets. For example, the NSSMF in CN may calculate the number of VNFs associated with computing, storage, and networking resources (i.e., VM's resources) according to the requirement of the network service received from NSMF. Then, the essential VNFs are initialized by NFV-MANO, as shown in the left side of Fig. 1. Resource allocation relies on the NSSMF with NFV-MANO that can calculate the required resources according to the requirement of network services. Moreover, NSSMF with NFV-MANO may dynamically adjust resource allocation by means of upgrading or scaling in/out network functions, which are triggered by the traffic fluctuation or the variation of network slice requirements. Under the MO architecture of network slicing, isolation levels and mathematical models play significant roles in resource allocation, which are discussed below.

Resource Type and Isolation

RAN Slicing

In general, the frequency spectrum is the fundamental radio resource for RAN slicing. In long term evolution (LTE) systems, the frequency spectrum resource is extracted as resource blocks (RBs) or physical resource blocks (PRBs). A network slice with dedicated PRBs strictly ensures the required QoS/SLA, while the common medium access control (MAC) scheduler can allocate and manage shared PRBs to accommodate elastic traffic, varying channel conditions and QoS requirements, thereby enhancing resource resilience and multiplexing gain. Other fundamental physical resources include transmission power and the cache space of the base station (BS), which are discussed later.

Based on PRB isolation, packet scheduling, as a high-level isolation, decides the timing to handle data traffic using available PRBs. The highest level of resource isolation is the admission control that determines whether to establish a data delivery service between users and the CN (i.e., radio access bearer) by estimating the state of network resource usage. More generally, admission control also includes to determine whether a user is allowed to access the network or whether a network slice request is accepted [6]. In addition, user association is a significant step of resource allocation in RAN slicing, which determines whether a user is associated with a specific BS. User association occurs before data transmission and after user admission control. It is difficult to identify a distinct isolation level for user association, but it is worth noting its crucial role for load balancing, radio spectrum efficiency, and network efficiency.

Compared with the traditional architecture of RAN, the baseband process and remote radio access are separated in Cloud-RAN (C-RAN). A centralized baseband unit (BBU) pool, distributed multiple remote radio heads (RRHs), a fronthaul network connecting the BBU pool and RRHs via optical links, and a backhaul network connecting the BBU pool to CN, are the four essential parts of C-RAN. The BBU pool is for baseband processing built on high-performance general-purpose hardware with virtualization techniques, which can run on VMs or containers to improve the utilization ratio of physical resources (e.g., computing, storage, networking resources). The distributed RRH provides radio access for users by antennas to cover a large area with low CAPEX and OPEX due to the low deployment cost. The capacities of the BBU pool and the RRHs with the fronthaul bandwidth are non-trivial factors when designing resource allocation algorithms. In addition, user association and backhaul capacity also affect resource allocation. We summarize the isolation levels associated with resource types for RAN slicing in the left side of Fig. 2.

CN Slicing

Compared with the EPC in 4G, the 5G core network adopts a more modular architecture by dividing networks into more fine-granular network functions [4]. Each network function can be deployed on a virtual platform in the manner of a VM or container, which can be managed and orchestrated by SDN and NFV to provide flexible, scalable, and programmable network services. A high level of resource isolation in CN slicing can be the VM or container, which means that network operators exploit the operation of scaling in/out at the VM or container granularity to initialize or to adjust resource allocation.

Under the framework of NFV-MANO, a CN slice, comprised of multiple VNFs, forms a virtual network running on a substrate network. In this

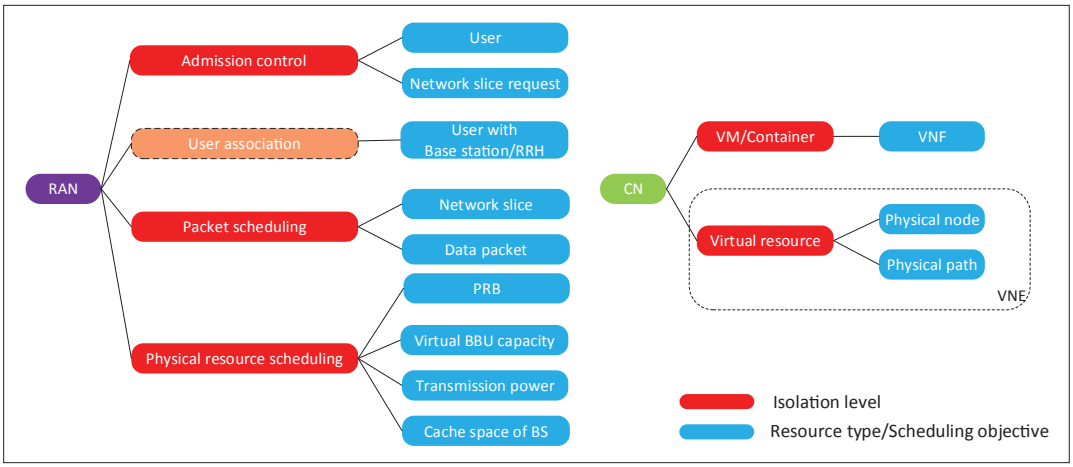

FIGURE 2. Isolation levels with resource types of network slicing.

case, the resource allocation of a CN slice is converted to the virtual network embedding (VNE) problem. The VNE poses two issues: mapping virtual nodes to physical nodes, and mapping virtual links connecting virtual nodes to paths connecting physical nodes. The physical nodes and paths represent computing, storage, and networking resources. In our opinion, the VNE can be utilized to model the resource allocation problems in CN slicing as well as some parts of C-RAN slicing (e.g., virtualization of the BBU's capacity). Comprehensive surveys regarding the VNE problem can be found in [4]. To avoid repetition, we summarize and analyze the latest typical mathematical models for 5G CN slicing later. The right side of Fig. 2 shows the isolation levels and resource types in CN slicing. In Fig. 3 we also describe the relationship among radio resource management, the VNE, and the resource allocation of network slicing.

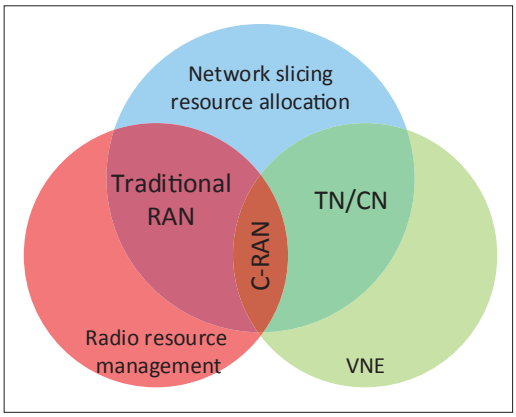

FIGURE 3. Relationship among resource allocation of network slicing, radio resource management, and VNE.

Mathematical Models

General Models

It is straightforward to formulate the resource allocation problem of network slicing as a linear programming (LP) problem or nonlinear programming (NLP) problem. The optimization objectives include the throughput of network slices, the resource utilization ratio, the remaining physical resources for the next assignment, and so on. The general constraints are the transmission power of a BS, the minimum number of PRBs to each network slice based on service contracts, the allocation fairness among different network slices, etc., which may become complicated due to varying network environments and diverse requirements of network slices. Thus, some general optimization problems are difficult to be solved in polynomial time. Heuristic approaches incorporated with multiple isolation levels are employed to derive near-optimal solutions with low computational complexity.

Furthermore, as discussed previously, C-RAN centralizes the baseband process capacity by means of the BBU pool that connects to several RRHs via fronthaul networks and to CNs via backhaul networks. Therefore, the capacities of BBU pools and the fronthaul/backhaul networks, as new constraints, are adopted in the framework of general optimization problems for C-RAN slicing. For example, the authors in [7] propose a hierarchical resource allocation scheme to maximize the network throughput of all network slices in C-RAN [7]. The hierarchical scheme includes admission control, user association, virtual BBU capacity allocation as an upper-level allocation and the PRB and transmission power allocation for each user in the corresponding network slice as a lower-level allocation. The computational burden and large signaling overhead are significantly alleviated when the optimal results from the upper-level allocation are utilized in the lower-level allocation.

Moreover, the resource allocation for end-to-end network slices is also explored by leveraging the general mathematical model. The resources in RAN, transport network (TN), and CN, are abstracted as the computing and the bandwidth resources [8]. Then, the resource allocation problem is converted to the VNE problem, which has different objectives according to different network slice types. For example, the optimization objective is to minimize the packet delay for the ultra-reliable and low-latency communications (uRLLC) network slice. The number of remaining physical links after current network slice deployment should be maximized when assigning resources for the massive machine type communications (mMTC) network slice.

Price differentiation is a simple approach based on the economic and game model, which assigns different prices for different network slices according to their characteristics, such as traffic model, resource capacity, required resource quantity, and network slice priority. The revenue of infrastructure providers is a typical optimization objective.

Game-Theoretic Economic Models

Network slicing provides new business opportunities for traditional network operators, vertical industries, and third parties. There are three typical players in the scenario of 5G network slicing: the network operator, the tenant of a network slice, and the user of a network slice. The network operator (e.g., infrastructure provider) provides the network infrastructure including physical resources and virtual resources and establishes the network slice to meet the requirements of network services. The tenant of a network slice (i.e., network service provider) proposes the network service requirement to the network operator and provides the network service for the users. The user (e.g., user equipment (UE)) uses the network service provided by the network slice. Compared with the general models discussed above, the economic and game models offer sophisticated and efficient approaches to describe the relationship among these three players, which can formulate the optimization problems, such as the fairness of network resource allocation, profit maximization of network operator or tenant of a network slice, and cost minimization of a network slice's users. For example, the economic factors, including budget, cost, revenue, profit, and price, significantly affect the resource allocation during the game among these players. The economic and game models can provide feasible solutions of the revenue or profit maximization of infrastructure providers without sacrificing the fairness of resource allocation.

Price differentiation is a simple approach based on the economic and game model, which assigns different prices for different network slices according to their characteristics, such as traffic model, resource capacity, required resource quantity, and network slice priority. The revenue of infrastructure providers is a typical optimization objective. For example, the network slices are priced differently according to their traffic models, including elastic and inelastic data traffic models [6]. Moreover, a relationship between the numbers of the elastic network slices and inelastic network slices is derived, which is under the restrictions of the maximum number of admitted inelastic network slices and the fixed sum of inelastic and elastic network slices. Based on the relationship and the price differentiation, a value iteration algorithm incorporated with semi-Markov decision process (SMDP) is proposed to search an optimal admission control of network slice requests to maximize the revenue of the infrastructure providers [6].

Fisher market, as a well-known economic model, is utilized to investigate the fairness among users and networks slices [9]. The buyers with fixed budgets in the Fisher market are mapped to the network slice tenants with shared network resources. The bid of a buyer in Fisher market is mapped to the preference of a network slice tenant. In the scenario of 5G network slicing, a tenant of a network slice may dynamically adjust its preferences depending on resource congestion to maximize its own utility. In this case, the resource allocation problem is converted to how to allocate resources for network slices by jointly considering their preferences and users' fairness through a non-cooperative game. Unlike predicting and analyzing the behavior of each player in the non-cooperative game, the cooperative game investigates the alliance among players formed by individual players or external enforcement (e.g., contract rules). The compositions of an alliance, the joint actions when generating an alliance, and the collective payoff when an alliance has already formed are the research interests of the cooperative game. Take CN slicing as an example. The VNF placement problem is solved under the framework of the cooperative game by two steps [10]. First, in the same cloud network, an optimization problem is formulated aiming at minimizing the creation cost of VNFs by tuning the number of VNFs. Then, how to place these VNFs across different cloud networks is discussed, which guarantees the QoS/SLA and maximizes the profit of each cloud network. In fact, different cloud networks with deployed VNFs can form different coalitions. A merge and split rule is proposed to analyze profits of all possible coalitions.

Prediction Models

Prediction models are utilized to forecast the proper or optimal resource quantity based on empirical knowledge or historic information. The predicted objects include arrival rate of users, data traffic in a network slice, network slice requests, and so on. Simple prediction approaches directly use empirical data or some well-known probability models. For example, different from VNFs placement using the cooperative game [10], the positions of VNFs are determined by the probability distribution of the arrival rate of session requests. [11]. The arrival rate of session requests is assumed to follow non-uniform distribution in the physical domain, which is converted to uniform distribution in the canonical domain by the Schwarz-Christoffel mapping. Then, the optimal number and positions of VNFs in the canonical domain are easy to derive according to the volume of session requests and the arrival rate of session requests. The corresponding positions of VNFs in the physical domain can be obtained by the inverse spatial transformation function.

Furthermore, the time-series based prediction model, as a regression prediction, forecasts the statistical characteristics in network slices based on past observations along with time series. Compared with empirical knowledge, the time-series based prediction model is more practical and suitable for varying network environments. Better than the basic and double exponential smoothing models, the Holt-Winters (HW) model, as a sophisticated time-series model, can capture not only the data trend, but also the data seasonality. In [12], the HW model is employed to analyze future traffic requests in network slices when the number of users is fixed. Furthermore, the traffic model of network slices when users moving in a multi-cellular environment is also predicted by the HW model. It is worth noting that a feedback mechanism is designed for each admitted network slice to improve the prediction accuracy of the HW model [12].

Machine learning is adopted to generate an optimal or suboptimal strategy based on the historic data and decisions. It is suitable for the scenario without a unified probability distribution. For instance, support vector regression (SVR), as a typical classification algorithm in machine learning, is employed to predict the values of the statistical characteristics of network slices in a time division duplexing (FDD)-OFDMA wireless network [13]. To be more specific, the authors in [13] realize the resource allocation of RAN slicing for both mobile edge computing (MEC) and traditional network services for different configurations of uplink/downlink. The configuration is formulated as a closed-form expression with the statistical characteristics of a network slice, including energy overhead, run-time overhead, and traffic and data rate of uplink and downlink. The optimal configurations for MEC and traditional network services are derived from the closed-form expression with the values of the statistical characteristics predicted by SVG. Moreover, Q-learning is another efficient candidate for prediction, which is utilized to make admission control decisions of network slice requests with low computational complexity [6].

Robustness and Failure Recovery Models

The resource allocation algorithms of network slicing should not only enhance the resource utilization efficiency, but also handle unpredictable network events to achieve a high availability of telecommunication networks. Unpredictable network events include network congestion (caused by heavy data traffic) or network function failure (caused by unexpected malfunction of software or hardware). Redundant resource reservation and network function remapping are two efficient approaches.

Redundant resource reservation offers extra resources for network slices, such as multiple instances of VNF, to avoid network function failure. The uncertain traffic of network slices is formulated as a symmetric and random variable by the chance-constrained model [14]. This variable, serving as a protection level, is added into the capacity constraints of virtual nodes and virtual links respectively to guarantee network slice robustness. The remapping of network functions recovers network services by mapping failed VNFs to good ones and rerouting failed virtual links. The optimization objective is to minimize the total bandwidth consumption during remapping [15].

Summary

In Fig. 4, we illustrate the relationship between the four mathematical models. The proportion of each approach in every pie chart represents the corresponding research potential. We also summarize the discussed algorithms in terms of objectives, resource type, scenario, and key descriptions in Table 1.

Future Directions and Challenges

Update Period of Resource Allocation

As discussed previously, the computational complexity of resource allocation algorithms should be reduced as much as possible. Therefore, heuristic approaches are proposed to provide suboptimal solutions with low computational complexity. Resource allocation by multiple levels is another approach, which includes admission control, user association, and so on. However, few research reports or publications discuss the update period of the resource allocation algorithm. An appropriate update interval not only improves the resource utilization ratio, but also decreases the signaling overhead and the computational burden. In [7], the authors point out that the period of allocating PRBs cannot be shorter than that of the channel condition reporting. A longer update period is a feasible solution in slow-varying channels, but may not achieve high performance in fast-varying channels. Moreover, the variation of the number of users, the fluctuation of data traffic, and the mean cycle between failures of network functions, can be jointly considered when designing the update period of resource allocation. We believe that there is still scope for this direction.

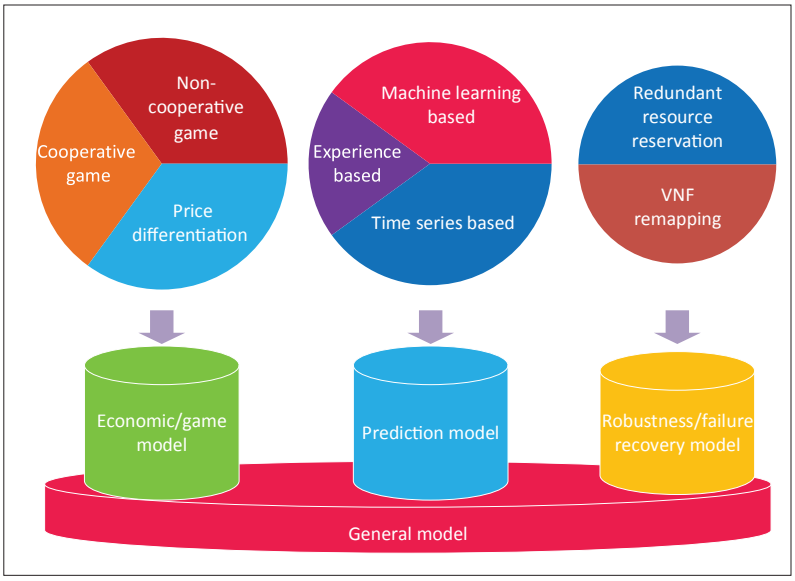

FIGURE 4. Relationship among different mathematical models with corresponding approaches.

Heterogeneity of Radio Access Technologies (RATs)

5G is expected to span and aggregate multiple heterogenous RATs, such as 3G, 4G, LTE, and WiFi. The cooperation of different RATs becomes significant in 5G network slicing, which should meet the diverse demands of network slices. For example, the vehicular infotainment application in a vehicle to everything (V2X) network slice relies on LTE and WiFi to improve the network throughput. The narrowband IoT (NB-IoT) application running on a mMTC network slice may require diverse RATs to enhance the network connectivity. Current algorithms focus on dense heterogeneous wireless networks and investigate the cooperation between macro-cells and small-cells [7]. For different network slices, resource allocation with multiple RATs to enhance the user experience and improve the resource utilization ratio can be further studied in terms of seamless handover, traffic isolation, and network slice fairness.

Resource Allocation for End-to-End Network Slicing

Most current algorithms of resource allocation focus on a single subnet, such as RAN or CN. Besides the solutions in [8], few researchers consider the end-to-end network slice. In [8],

Ref.	Math model	Main objectives	Resource type	Scenario	Key description
Lee [7]	General model	Throughput of network slices	PRB, BBU capacity, transmission power	C-RAN	A hierarchical resource allocation: admission control, user association, virtual BBU capacity, PRB/transmission power allocation
Guan [8]		Packet delay, remaining resource	Computing, bandwidth resource	RAN + CN	Resource allocation formulation by VNE/complex network theory to obtain network topologies
Bega [6]	Economic/game model	Revenue of infrastructure providers	Network slice	RAN	Price differentiation for different traffic models/Q-learning: optimal decisions on accepting/rejecting network slice request
Caballero [9]		Data rate of all users	PRB	RAN	Fisher market and a fair utility function: fairness/more stability and substantial gains from dynamic network slicing than that of static network slicing
Bagaa [10]		Cost of creating VNFs	VNF	CN	Optimal number of VMs for each VNF/VNFs on different cloud networks forming coalition to minimize the creation cost of VNFs
Laghrissi [11]	Prediction model	Packet delay and total number of VNFs	VNF	CN	Arrival rate of session requests transform between non-uniform and uniform by Schwarz-Christoffel mapping
Sciancalepore [12]		Resource utilization ratio	PRB network slice	RAN	HW model: predict trend and seasonal feature of future data traffic
Zhao [13]		Energy consumption + run-time overhead	PRB	RAN	SVR: predict statistical features of network slices/ achieve resource isolation among network slices and service customization in each network slice
Baumgartner [14]	Robustness/failure recovery model	Capacity consumption installation cost of remap	VNF	CN	Uncertain traffic: and random variable/reduce redundant resources: survivability+robust symmetric
Wen [15]		Total bandwidth of all virtual links	VNF	CN	Remap failed VNFs to new ones/rerouting failed virtual links by using VNE

TABLE 1. Comparison of resource allocation algorithms for network slicing in 5G.

the resources of the radio and VM are abstracted as bandwidth and computing resources, respectively. Only one of them is considered when designing a resource allocation scheme for a particular network slice type. In fact, SLA decomposition and multiple subnet coordination are the main challenges of resource allocation for an end-to-end network slice. The tenant of a network slice may only provide an end-to-end SLA to NSMF without the requirement for each subnet due to a lack of fundamental communication knowledge. Therefore, how to decompose the end-to-end SLA into each subnet's requirement is an inevitable step in resource allocation. SLA decomposition should consider not only the end-to-end capacity, but also the capacity of each subnet. The coordination among multiple subnets is also significant especially when updating resource allocation. Each subnet should report its remaining capacity to NSMF periodically via NSSMF and NSMF can achieve an appropriate adjustment of resource allocation to accommodate the varying network environment. For instance, if some subnets are overloaded, they should report to NSMF. Then, the remaining subnets are allowed to obtain more resources to guarantee the desired end-to-end SLA. Potential mathematical models should be carefully designed based on the multi-operator core network (MOCN) and gateway core network (GWCN), which are two fundamental architectures of resource sharing among different subnets [3].

Conclusions

This article surveys the resource allocation schemes for 5G network slicing in terms of principles and mathematical models. As the key enabling techniques, SDN and NFV are introduced with the MO architecture of network slicing. The resource types with corresponding isolation levels in RAN and CN are analyzed. We categorize mathematical models of resource allocation into four types according to their objectives: general, economic and game, prediction, and robustness and failure recovery models. The motivation, objective, and main principle of each model are elaborated and analyzed with the latest examples. In addition, the envisioned future directions in line with the latest trend of end-to-end network slicing in 5G networks are presented.

References

[1] J. W. L. Zhao, J. Liu, and N. Kato, "Routing for Crowd Management in Smart Cities: A Deep Reinforcement Learning Perspective," *IEEE Commun. Mag.*, 2019, pp. 1–1.

[2] P. Rost *et al.*, "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks," *IEEE Commun. Mag.*, vol. 55, no. 5, May 2017, pp. 72–79.

[3] S. Abdelwahab *et al.*, "Network Function Virtualization in 5G," *IEEE Commun. Mag.*, vol. 54, no. 4, Apr. 2016, pp. 84–91.

[4] I. Afolabi *et al.*, "Network Slicing & Softwarization: A Survey on Principles, Enabling Technologies & Solutions," *IEEE Commun. Surveys & Tutorials*, 2018, pp. 1–24.
[5] S. Vassilaras *et al.*, "The Algorithmic Aspects of Network Slicing," *IEEE Commun. Mag.*, vol. 55, no. 8, Aug. 2017, pp. 112–19.
[6] D. Bega *et al.*, "Optimising 5G Infrastructure Markets: The Business of Network Slicing," *Proc. IEEE Conf. Computer Commun. (INFOCOM)*, Atlanta, GA, May 2017, pp. 1–9.
[7] Y. L. Lee *et al.*, "Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, Apr. 2018, pp. 2146–61.
[8] W. Guan *et al.*, "A Service-Oriented Deployment Policy of End-to-End Network Slicing Based on Complex Network Theory," *IEEE Access*, vol. 6, Apr. 2018, pp. 19 691–01.
[9] P. Caballero *et al.*, "Network Slicing Games: Enabling Customization in Multi-Tenant Networks," *Proc. IEEE Conf. Computer Commun. (INFOCOM)*, Atlanta, GA, May 2017, pp. 1–9.
[10] M. Bagaa *et al.*, "Coalitional Game for the Creation of Efficient Virtual Core Network Slices in 5G Mobile Systems," *IEEE JSAC*, vol. 36, no. 3, Mar. 2018, pp. 469–84.
[11] A. Laghrissi, T. Taleb, and M. Bagaa, "Conformal Mapping for Optimal Network Slice Planning Based on Canonical Domains," *IEEE JSAC*, vol. 36, no. 3, Mar. 2018, pp. 519–28.
[12] V. Sciancalepore *et al.*, "Mobile Traffic Forecasting for Maximizing 5G Network Slicing Resource Utilization," *Proc. IEEE Conf. Computer Commun. (INFOCOM)*, Atlanta, GA, May 2017, pp. 1–9.
[13] P. Zhao *et al.*, "Information Prediction and Dynamic Programming Based RAN Slicing for Mobile Edge Computing," *IEEE Wireless Commun. Lett.*, Feb. 2018, pp. 1–4.
[14] A. Baumgartner *et al.*, "Optimisation Models for Robust and Survivable Network Slice Design: A Comparative Analysis," *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–7.
[15] R. Wen *et al.*, "Protocol Function Block Mapping of Software Defined Protocol for 5G Mobile Networks," *IEEE Trans. Mobile Computing*, vol. 17, no. 7, July 2018, pp. 1651–65.

Biograpahies

Ruoyu Su received his B.Eng. and M.Eng. degrees from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2006 and 2009, respectively, and his Ph.D. degree in electrical and computer engineering from Memorial University, St. John's, NF, Canada, in 2015. He completed his postdoctoral work in 2016 in electrical and computer engineering at Memorial University. He is currently a faculty member with Nanjing University of Post and Telecomm. His research interests include energy-efficient design for underwater acoustic sensor networks, cross-layer design for wireless sensor networks, and 5G network slicing.

Dengyin Zhang received the B.S. degree, M.S. degree, and Ph.D. from Nanjing University of Posts and Telecommunication, Nanjing, China, in 1986, 1989, and 2004, respectively. He is currently a professor and also serves the Dean of the School of Internet of Things, Nanjing University of Posts and Telecommunication, Nanjing, China. He was at the Digital Media Lab at Umea University in Sweden as a visiting scholar from 2007 to 2008. His research interests include signal and information processing, networking techniques, and information security.

Ramachandran Venkatesan [M'78, SM'92] received the B.E. degree (Hons.) from Madurai University, Madurai, India, and the M.Sc.E. and Ph.D. degrees from the University of New Brunswick, Fredericton, NB, Canada, all in electrical engineering. He is a professor of computer engineering with Memorial University of Newfoundland, St. John's, NL, Canada, where he has been working since 1987. He worked in Industry as a welding research engineer for several years. He has held several academic administrative positions including the Chair of Electrical and Computer Engineering, the Associate Dean for Graduate Studies and Research, the Associate Dean for Undergraduate Studies, the Acting Dean, and the Dean Pro Tempore of Engineering. His research interests include parallel processing architectures and applications, error control coding, underwater communications, wireless communications, and optical communications. He is a Registered Professional Engineer.

Zijun Gong received the B.Eng. and M.Eng. degrees from Harbin Institute of Technology (HIT), Harbin, P. R. China, in 2013 and 2015, respectively. He has been pursuing his Ph.D. degree since then at Memorial University of Newfoundland (MUN), St. John's, NL, Canada. His research interests include WLAN fingerprint based indoor localization systems, radio propagation modeling, localization of WSN, and localization of underwater vehicles. Also, he did research on channel estimation in massive MIMO and millimeter wave communications. He is a reviewer for *IEEE Transactions on Vehicular Technology* and *IEEE Transactions on Communications*. He received the Best Paper Award at the 2017 IEEE Global Telecommunications Conference (Globecom'17), Singapore, December 2017.

Cheng Li [M'04, SM'07] received the B.Eng. and M.Eng. degrees from the Harbin Institute of Technology, Harbin, China, and the Ph.D. degree in electrical and computer engineering from Memorial University, St. John's, NF, Canada, in 1992, 1995, and 2004, respectively. He is currently a full professor in the Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Memorial University. His research interests include mobile ad hoc and wireless sensor networks, wireless communications and mobile computing, switching and routing, and broadband communication networks. He received the Best Paper Award at ADHOCNETS'18, IEEE Globecom'17, and IEEE ICC'10. He is an Editorial Board member of *China Communications*, *Wireless Communications and Mobile Computing*, an associate editor of *Security and Communication Networks*, and an Editorial Board member of the *Journal of Networks* and *KSII Transactions on Internet and Information Systems*. He has served as the General Co-Chair for AICON'19, WINCOM'19, and WINCOM'17, and he has served as a TPC Co-Chair for the ICNC'19, WiCON'17, MSWIM'14, MSWIM'13, and QBSC'10. He has also served as the Co-Chair for various technical symposia or tracks of many international conferences, including IEEE ICC, Globecom, IWCMC, VTC, and WCNC. He is currently the Elected Chair of the IEEE Communications Society Ad Hoc and Sensor Networks Technical Committee (AHSN TC). He is a Registered Professional Engineer in Canada and a senior member of the IEEE and its Communications, Computer, Ocean Engineering, and Vehicular Technology Societies.

Fei Ding received the Ph.D. degree in instrument science and technology from the School of Instrument Science and Engineering, Southeast University, Nanjing, China, in 2010. He was an Internet of Things (IoT) research leader with the R&D Center, China Mobile Group Jiangsu Co., Ltd., Nanjing China, and also a postdoctoral researcher in the School of Information Science and Engineering, Southeast University, Nanjing, China. He is currently an associate professor with the School of IoT, Nanjing University of Posts and Telecommunications, Nanjing, China. He has long been engaged in wireless sensor networks, IoT, and mobile communication related key technologies. He has chaired or participated in more than 10 National or Provincial Science and Technology Projects, and chaired over 20 Enterprise Projects.

Fan Jiang received the B. E. degree in communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, the M.E. degree in communication and information systems from Southeast University, Nanjing, China, and the Ph.D. degree in computer engineering from Memorial University of Newfoundland (MUN), St. John's, NL, Canada, in 2010, 2013, and 2018, respectively. His research interest includes localization of wireless sensor networks, and signal processing for wireless and underwater acoustic communication systems. He received the prestigious Governor General's Gold Medal at MUN in 2018. He was also the recipient of the Best Paper Award at the 2017 IEEE Global Telecommunications Conference (Globecom'17), Singapore, December 2017.

Ziyang Zhu is a visiting student at Memorial University in 2018. He studies at the University of Sheffield, UK.