Hierarchical Slicing: A New Paradigm of Radio Resource Management for Mobile Networks

Tianyu Wang, Xun Cao, and Shaowei Wang

Abstract

Radio access network slicing is considered to be highly challenging due to the complexity of mobile environments and the diversity of mobile services. Existing works in this area mainly decompose the radio resource management problem into slice-level resource management and user-level resource scheduling. In this paper, we propose a hierarchical three-tier slicing architecture, where an additional tier is introduced to merge the gap between the conventional two tiers in time and spatial scales. Specifically, the medium-scale tier provides inter-cell and inter-slice coordination to address short-term dynamics caused by local user mobility and traffic variations, which highly simplifies the upper and lower tiers by allowing them to focus on long-term traffic distribution and instant user demands, respectively. Thus, it allows each independent iter to apply more flexible and more efficient radio resource management methods with appropriate time and spatial scales. A proof of concept is provided to show that the proposed three-tier architecture can achieve a better tradeoff between slice isolation and slice capacity as compared to the state-of-the-art two-tier solution.

Introduction

Future mobile networks are required to support a variety of vertical applications, including the conventional 5G use cases, i.e., enhanced mobile broadband (eMBB) focusing on the multimedia traffic for human users, ultra-reliable low-latency communication (URLLC) for life-critical applications, and massive machine-type (mMTC) for Internet of Things, as well as the new additions for 5.5G, including machine vision for automatic inspection and security, virtual reality, and wide-area high-resolution sensing in self-driving and elder care. Due to the diverse quality of service (QoS) requirements of future mobile services, the conventional one-size-fits-all solution becomes unacceptable in both technical and economic aspects. Instead, network slicing is proposed as a novel solution for network architecture, which slices the physical network into multiple virtual networks, referred to as network slices, such that each slice can be customized for a specific service [1].

Network slicing is an end-to-end technology, which requires management and orchestration across all three domains of the mobile communication system, i.e. the radio access network (RAN), the transport network and the core network [2]. Among these domains, RAN slicing is regarded as the most challenging one. The major challenges are as follows.

- Radio resources in RANs are highly limited as compared to fiber resources in transport and core networks. Thus, RAN slicing raises higher demands in slicing efficiency as compared to network slicing in other domains.
- Real-time computing resources are also very limited in RANs. While the RAN slicing decisions need to be frequently adjusted due to the dynamics of wireless channels and local user demands. Thus, RAN slicing methods are required to satisfy stringent constraints in computational complexity.
- On the one hand, inter-cell inter-slice coordination is required to avoid unexpected interference and guarantee the slice performance. On the other hand, slices are logically independent networks that should be isolated from each other. Thus, the RAN slicing architecture must be able to achieve a tradeoff between performance guarantee and slice isolation, which becomes a challenging issue in multi-cell multi-slice RAN slicing.

As shown in Fig. 1, existing works of RRM in multi-cell multi-slice networks follow a two-tier architecture [3], [4]. In the upper tier, the available bandwidth of each cell is assigned to slices in a quasi-static manner according to their long-term traffic demands. In the lower tier, the dedicated radio resources of each slice are scheduled among the corresponding users according to the real-time dynamics. However, the complex interference coordination between different cells and slices poses huge technical challenges to the traditional two-tier networks, especially considering the dynamic environment of mobile networks and the diversity of slice requirements.

In [5], a low-complexity method is proposed for the radio resource allocation in upper tier, which greatly reduces the inter-slice interference as compared to the typical priority-based method.

Tianyu Wang (corresponding author) is with the School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; Xun Cao and Shaowei Wang are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China.

Digital Object Identifier:
10.1109/MNET.128.2200304
Date of Current Version:
10 May 2024
Date of Publication:
29 May 2023

0890-8044/23©2023IEEE

The proposed method shows superior performance in terms of the average throughput, while at the same time, suffers from high computational complexity in large-scale networks.

However, due to the heuristic steps of the proposed method, there exists no performance guarantee in a general case. In [6], a 0-1 quadratic programming based method is proposed for the upper tier, which aims to enhance the inter-cell transmission coordination and reduce the inter-slice interference at the same time by maximizing the number of "linked" radio resources. The proposed method shows superior performance in terms of the average throughput, while at the same time, suffers from high computational complexity in large-scale networks. Papa et al. [7] focus on the packet scheduling problem in the lower tier, for which a multi-cell multi-slice scheduler is proposed to jointly schedule radio resource blocks of neighboring cells and slices. However, due to the real-time requirement of packet scheduling and the diverse QoS requirements of network slices, the proposed method suffers from high computational complexity and may not be feasible in practical deployments.

In this paper, we consider the RRM issue in multi-cell multi-slice networks. Specifically, we first show that the fundamental difficulty of RRM in RAN slicing comes from the coupling effects between the cell and slice dimensions, and discuss the three key design aspects that always should be taken into account, i.e., slice capacity, slice isolation and implementation complexity. Then, we analyze the characteristics of network dynamics with different spacial and temporal scales, based on which we propose a hierarchical slicing architecture. As compared to the conventional two-tier architecture, the proposed architecture introduces an additional tier in the middle to adapt to the short-term variations caused by local user mobility and traffic patterns, which merges the huge gap between the quasi-static slicing and the real-time scheduling. At last, a proof of concept is provided to show that the proposed architecture outperforms the conventional architecture in terms of the isolation-capacity-complexity tradeoff.

RRM in Multi-Cell Multi-Slice Networks

In Fig. 2, we illustrate RRM in a network with two cells and three slices, where slices 1 and 3 provide eMBB services in both cells A and B, while slice 2 provides a URLLC service in cell B. As each slice is an independent virtual network with specific QoS requirements and coverage area, the corresponding RRM should involve not only the conventional cell dimension but also an additional slice dimension. Therefore, we have four different RRM issues in multi-cell multi-slice networks, which are discussed separately as follows.

- **Inter-Cell Intra-Slice Interference Coordination:** Any slice covering neighboring cells can generate inter-cell intra-slice interference. As shown in Fig. 2, two cell-edge users of slice 1 residing in cells A and B, respectively, are scheduled on a colliding resource block, which causes inter-cell interference of slice 1. To reduce the interference and improve the cell-edge performance, inter-cell intra-slice interference coordination can be performed by jointly coordinating the transmissions within a slice. We note that the slice-specific resource blocks are different for different cells due to the uneven distribution of slice traffic. Thus, the conventional inter-cell interference coordination methods, including flag-based reuse coordination in the frequency domain, almost blank subframes in the time domain, coordinated multipoint transmission in the spatial domain, and various power adjustment techniques should be modified before they can be applied in inter-cell intra-slice interference coordination.

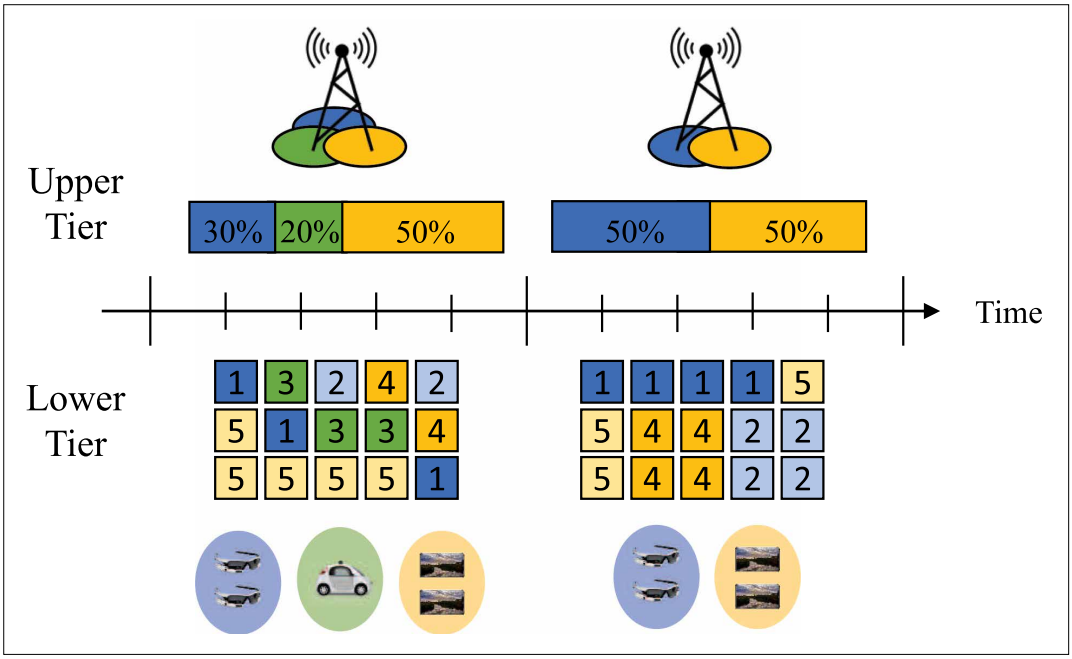

FIGURE 1. The conventional two-tier slicing architecture for RRM in multi-cell multi-slice networks.

- **Inter-Cell Inter-Slice Interference Coordination:** Slices with overlapping coverage areas can generate inter-cell inter-slice interference at their common cell boarders. As shown in Fig. 2, the cell-edge users of slice 1 in cell A and slice 2 in cell B are scheduled on a colliding resource block, which causes interference between slices 1 and 2 at the boarder between cells A and B. Inter-cell inter-slice interference not only degrades the performance of cell-edge users, but also jeopardizes the isolation between the corresponding slices. To protect slice isolation from inter-slice interference, inter-cell inter-slice interference coordination is recently proposed and draws a lot of attentions [6]. It is shown that inter-cell inter-slice interference coordination requires global optimization across all cells and slices, which is NP-hard in general.
- **Intra-Cell Intra-Slice Resource Allocation:** Dedicated radio resources of each slice within each cell should be dynamically assigned to the corresponding users, such that the QoS requirements of each user can be satisfied in terms of throughput, latency and priority. As shown in Fig. 2, the dedicated resource blocks of slice 3 in cell A are equally assigned to two users to support their video streaming traffic. We note that there exists many packet scheduling schemes that can be inherited from the conventional networks, e.g., proportional fair schemes for eMBB slices to balance user fairness and spectral efficiency, delay-oriented schemes for URLLC slices to guarantee packet delay, and semi-persistent schemes for mMTC slices to reduce signaling overhead and power consumption. However, due to the existence of multiple slices, the amount of slice-specific radio resources can be time-varying and these scheduling schemes should be reconsidered before they can be applied.
- **Intra-Cell Inter-Slice Resource Allocation:** To provide QoS guarantee on a per slice basis, radio resources should be dynamically allocated among slices within a cell according to their time-varying traffic demands and channel conditions, which is referred to as intra-cell inter-slice resource allocation. As shown in Fig. 2, the resource blocks of cell B are equally assigned to slices 1, 2 and 3 to fulfill the QoS requirements of virtual reality, vehicular communications and video streaming, respectively. Due to the online operation and the incomplete feedback information, intra-cell inter-slice resource allocation is regarded to be highly challenging, for which both model-based and data-based methods have been proposed. Specifically, an online learning framework that is able to combine both the model-based optimization techniques and the data-based machine learning techniques has shown great advantages in providing efficient and robust intra-cell inter-slice resource allocation [8].

The additional slice dimension requires to consider not only the user-level performance but also the slice-level performance. The traditional methods cannot guarantee sufficient radio resources in the slice level or address the additional limitations introduced by slice isolation. Thus, in multi-slice multi-cell networks, the above RRM issues are highly coupled with inconsistent or even conflicting objectives. For example, inter-cell inter-slice interference coordination may restrict the available radio resources of a specific slice due to inter-slice interference, while intra-cell inter-slice resource allocation may require excessive radio resources for the slice to fulfill the QoS requirements. Therefore, it is important to reconsider the design principles and the appropriate framework of RRM in multi-cell multi-slice networks.

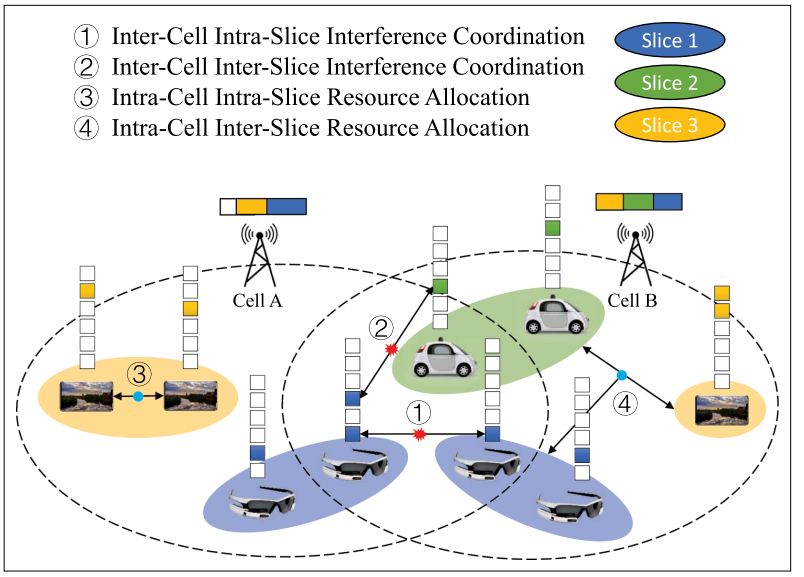

FIGURE 2. Illustration of RRM in a network with two cells and three slices. The available radio resources are represented by a column of blocks, where each color represents the part allocated to a specific slice or a specific user.

Key Design Aspects

There are three key design aspects that always should be taken into account for RRM in multi-cell multi-slice networks, i.e., slice isolation, slice capacity and implementation complexity. In this section, we will specify these aspects and discuss their effects on RAN slicing.

Slice Isolation

Slice isolation indicates the ability of an RRM solution to provide steady performance on each specific slice, regardless of the dynamics of other slices. For example, a URLLC slice needs to guarantee its average packet delay during the traffic burst of a coexisting eMBB slice. The RRM entity is required to simultaneously guarantee the performance of multiple slices, which are usually with highly different QoS requirements. Thus, the corresponding RRM strategy should be able to predict the performance variations of different slices for all possible traffic and channel changes, and make real-time decisions to ensure all performance indicators remain within their target ranges.

In the extreme case, the RRM strategy assigns to each slice an exclusive chunk of spectrum and an associated packet scheduler [9], which allows the slices to provide fully isolated services as if the slice users are served by a private network.

Slice capacity can be seen as an extension of the conventional system capacity, while it explicitly distinguishes between the traffic with different QoS requirements and thus is a more appropriate metric to evaluate RAN slicing strategies.

In the more general case, the RRM strategy allows the performance of a specific slice to be influenced by the dynamics of other slices, as long as such variations can be accurately controlled within an acceptable range. It presents novel challenges in terms of modeling and algorithms as compared to existing RRM strategies.

In addition, different slices may accept different levels of isolation due to both technical and business concerns, e.g., a live streaming slice may require a large amount of dedicated radio resources to maintain consistency of its user experience, while a web browsing slice may allow to flexibly share its radio resources with other slices since the browsing service is not sensitive to packet delay. Thus, the RRM strategy should be able to provide unified isolation levels, e.g., the isolation level of a specific type of slice can be defined as the maximal variations of its key performance indicators due to the presence of certain other slices [8]. Note that a slice may also experience performance variations due to its own dynamics, which however have no concern with slice isolation.

Slice Capacity

Slice capacity indicates the capability of an RRM strategy to support certain types of slices under a given amount of radio resources. Specifically, slice capacity is represented by a connected region in a multi-dimensional traffic space, where each dimension corresponds to a specific type of slice [8]. Any traffic combinations inside the region can be fulfilled, while any traffic combinations outside the region causes QoS violation. Thus, the RRM strategy should be able to expand the boundaries of the region as much as possible so as to maximize its slice capacity, while at the same time keep flexible to cover different traffic combinations in typical scenarios.

Slice capacity can be seen as an extension of the conventional system capacity, while it explicitly distinguishes between the traffic with different QoS requirements and thus is a more appropriate metric to evaluate RAN slicing strategies. However, due to the multi-dimensionality of traffic space, it is generally difficult to directly compare the slice capacity of different RRM strategies. For example, as compared to a proportional fair scheduler, a delay-oriented scheduler can support more URLLC traffic by providing bounded packet delay and less eMBB traffic due to the under-utilization of multi-user gain. In fact, an RRM strategy has a strictly larger slice capacity than another RRM strategy, only if its feasible traffic region fully covers the latter one. In practice, we can compare the slice capacity in serval predefined scenarios with typical traffic settings [10].

Implementation Complexity

Implementation complexity indicates the overall cost to deploy an RRM strategy in practical networks, involving algorithmic complexity, signaling overhead, synchronization and data collection requirements, which further affect the hardware cost, energy consumption and radio resource utilization. For example, QoS-aware schedulers require periodic weight calculation and channel state information feedback in each transmission interval, which highly increases the hardware cost and bandwidth consumption [11]. Joint transmission at cell edges relays on precise synchronization between neighboring cells, which requires wired backhaul with ideally large bandwidth and low latency [12]. Data-based schedulers usually adopt machine learning methods, which require a large amount of data for offline training and dedicated computational hardware for online inference [13]. Thus, the implementation complexity of RAN slicing strategies must be assessed carefully.

On the one hand, the additional slice dimension requires the strategy to deal with more complex RRM issues than conventional RRM schemes. On the other hand, due to the rapid evolution of cloud computing and wireless communication, the unit cost of general computation and communication capability keeps decreasing. Therefore, a relative high-complexity strategy that simultaneously fulfills multiple slices can still be cost-efficient. However, due to the high dynamics of multi-cell multi-slice networks, it is generally impossible to implement such a strategy with a single RRM entity. One possible solution is to apply a hierarchical structure to decompose RRM into different layers so as to simplify the RRM problem in each layer. However, such decomposition requires low coupling between layers and high efficiency within each layer.

Tradeoff

There exists intrinsic contradictions among the three design aspects discussed above. To achieve low implementation complexity and high slice isolation, static strategies can be applied, where dedicated radio resources are reserved for each slice. However, the required radio resources of static allocation are determined by the worst-case scenario, which highly degrades the average resource efficiency and results in low slice capacity. To increase the slice capacity, radio resources should be dynamically allocated according to the real-time traffic variations. The corresponding coupling effect between slices and cells may lead to high implementation complexity and low slice isolation. In fact, any RRM strategy can only guarantee its optimality in terms of two of the three aspects, which we refer to as the isolation-capacity-complexity tradeoff. We note that, for any RRM strategy with any isolation-capacity-complexity tradeoff, the QoS requirements should always be fulfilled without any compromise. Therefore, the design principle of RRM in multi-cell multi-slice networks is to guarantee the QoS requirements while achieving the best isolation-capacity-complexity tradeoff in interested scenarios.

Hierarchical Network Dynamics

Due to the complex mobility patterns of cellular users and the irregular fluctuations of slice traffic, RRM in multi-cell multi-slice networks must take account of the instant network dynamics and make online decisions to guarantee the QoS

requirements in such non-stationary environments. Therefore, an effective RRM strategy must identify the characteristics of different network dynamics. In this section, we will discuss in detail the network dynamics in a multi-cell multi-slice network and classify them into three different groups according to their spatial and temporal scales.

Large-Scale Dynamics

Large-scale dynamics refer to the long-term tendency of user mobility and traffic fluctuation in the granularity of cells and slices, e.g., the average number of users within a cell and the average amount of traffic of a specific slice. These large-scale dynamics vary slowly with time and can be regarded as quasi-static processes with a minute-level time window. Within each window, the large-scale parameters are static.

Large-scale dynamics have a decisive impact on the amount of radio resources required by the slices within a large geographical area. For example, the eMBB throughput at a subway station changes periodically as the train enters and leaves the station, while the URLLC throughput near a major road stays high during peak hours. We note that large-scale dynamics are highly related to the regularities in our social life [14], e.g., the travel patterns followed by different users, the service time of massive Internet of Things devices and the entertainment activities scheduled at different places. Thus, large-scale dynamics are steady and to a fair degree predictable, especially when historical data are available.

Medium-Scale Dynamics

Medium-scale dynamics refer to the short-term variations of network parameters caused by local user mobility and slice traffic patterns, e.g., the handover of cell-edge users and the on/off behaviors of slice-specific devices. As compared to the large-scale dynamics, medium-scale dynamics have a much shorter time window measured in seconds, during which the medium-scale behaviors are presumed to be unchanged.

As medium-scale dynamics represent the local behaviors of a small group of users, the corresponding effects on RRM are restricted to the corresponding slices and cells. For example, an eMBB user crossing the cell border changes the eMBB throughput in the serving and target cells, while the amount of radio resources required by other slices and other cells are not affected. As compared to the long-term tendency determined by large-scale dynamics, the short-term behaviors of medium-scale dynamics are in general more difficult to predict. However, if there exists a regular pattern or a logic model, machine learning methods based on deep neural networks can be applied to predict the time series data of medium-scale dynamics [15]. We note that the prediction accuracy is problem-dependent and may vary significantly.

Small-Scale Dynamics

Small-scale dynamics refer to the instant variations of the channel parameters and date rates of individual users, which reflect the inherent uncertainty of wireless environments and mobile applications in the granularity of milliseconds, e.g., the movement of interacting objects in the propagation environment and the real-time data rate of streaming media. Due to the randomness of such instant variations, small-scale dynamics are usually unpredictable and can only be characterized by using statistical models, e.g., the Rayleigh fading model and the Poisson process traffic model. In fact, small-scale dynamics are essentially non-stationary processes with time-varying model parameters, which are difficult to estimate and track in practical networks.

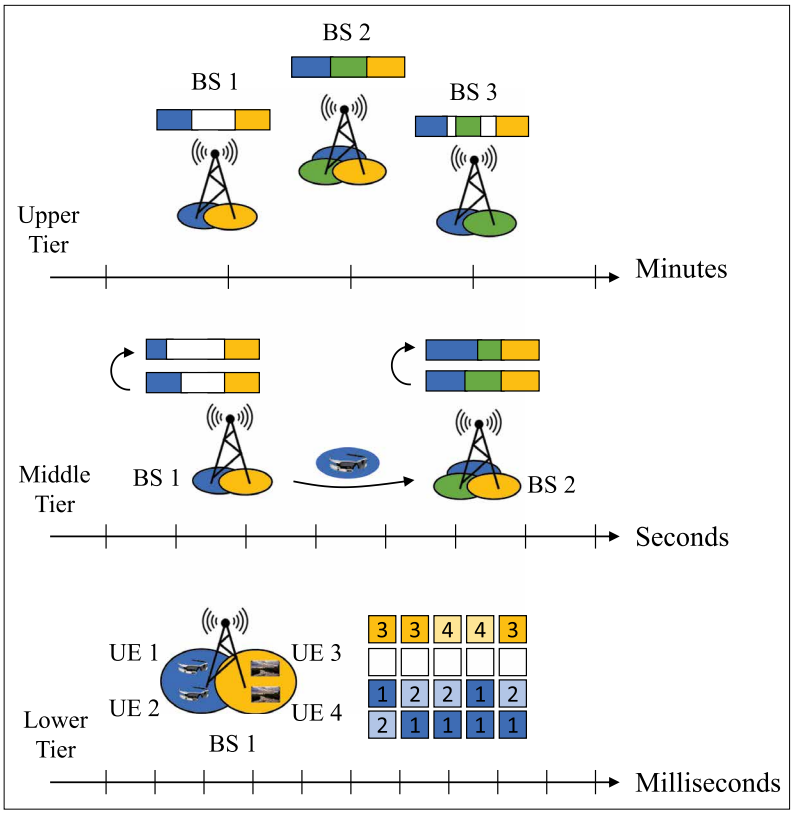

FIGURE 3. The hierarchical slicing architecture for RRM in multi-cell multi-slice networks.

Hierarchical Slicing Architecture

In this section, we propose a three-tier slicing architecture for RRM in multi-cell multi-slice networks, where each tier corresponds to a particular level of network dynamics with similar time and spatial scales. Thus, the proposed slicing architecture can divide and simplify multi-cell multi-slice RRM, while at the same time, ensuring a unified and efficient solution in each tier. In the upper tier, it is responsible for addressing the large-scales dynamics, for which the dedicated radio resources of each slice in each cell are reassigned across the entire network at the time granularity of minutes. In the middle tier, it is responsible for addressing the medium-scale dynamics, for which the local radio resource allocation is adjusted in a distributed manner at the time granularity of seconds. In the lower tier, it is responsible for addressing the small-scale dynamics, for which the instant radio resources allocated to each individual user is scheduled at the time granularity of milliseconds. The proposed architecture is shown in Fig. 3 and each tier is discussed in detail as follows.

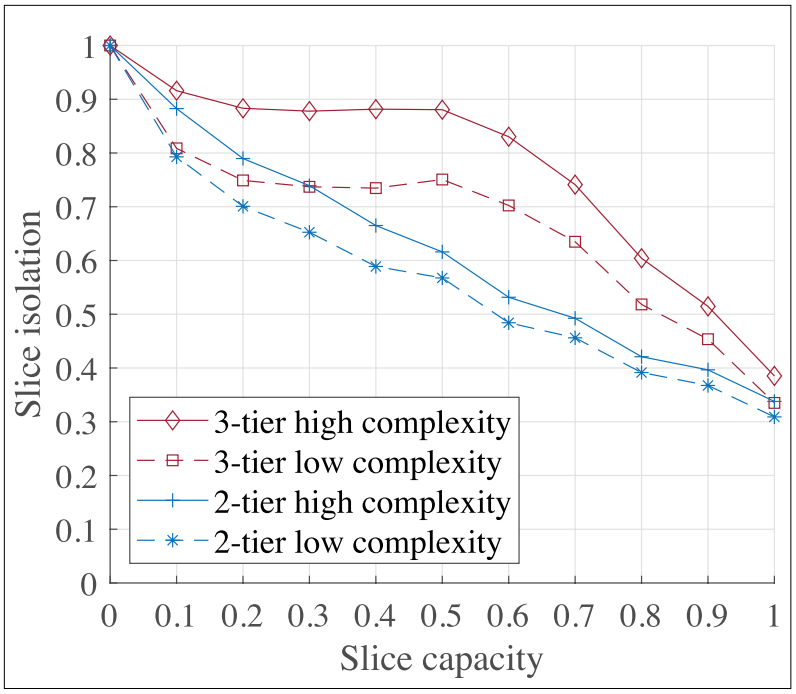

FIGURE 4. The isolation-capacity-complexity tradeoff of the proposed three-tier architecture and the conventional two-tier architecture for RRM in multi-cell multi-slice networks.

Radio Resource Assignment

In the upper tier, referred to as the radio resource assignment tier, the global radio resource allocation is periodically optimized to adapt to the large-scale network dynamics. Specifically, in each time window, the dedicated resource blocks of each slice in each cell is assigned according to the current network requirements. On the one hand, such network requirements may include coupling constraints in both the cell and slice dimensions, e.g., the amount of required radio resources of each slice in each cell, the coordination opportunities between cell-edge users of the same slice, and the interference level between different slices in neighboring cells. Thus, the optimal radio resource assignment falls into the field of integer programming, which is generally NP-hard and requires a large amount of computation time even with a medium network scale [6]. On the other hand, the static window of large-scale dynamics is bounded. Thus, the assignment period needs to be short enough to track the large-scale dynamics. Therefore, we need to choose a proper window size such that the RRM decisions can track the large-scale dynamics, while at the same time achieving high performance in every time window. In addition, if the large-scale variations can be detected or even predicted with high accuracy, the corresponding radio resource assignment can be event-triggered instead of following a periodic pattern.

Inter-Cell Inter-Slice Coordination

In the middle tier, referred to as the inter-cell inter-slice coordination tier, the radio resource allocation is adjusted locally in a distributed manner to adapt to the medium-scale network dynamics. Specifically, in each time window, the resource blocks are adjusted between neighboring slices and cells according to their requirement variations, e.g., the traffic demands of local slices, the coordination transmission requirement of cell-edge users, and the inter-slice interference bound of neighboring slices. Due to the volatile window of the medium-scale dynamics, this tier can no longer be regarded as a quasi-static process that consists of a sequence of static optimization problems, but an online optimization process that gradually learns the network dynamics from the information revealed by previous allocations [8]. If the relationship between the radio resource allocation and the target performance metric can be formulated by an analytical function, the online process is model-based, for which statistical gradient methods can be applied. If there exists no such model, the online process is data-based, for which deep learning methods can be applied. The network performance can be further improved if the medium-scale dynamics are predictable.

Packet Scheduling

In the lower tier, referred to as the packet scheduling tier, the dedicated resource blocks of each slice in each cell are scheduled in real time among the corresponding users, such that the user-level QoS requirements can be satisfied. The existing packet schedulers can be inherited. Note that the buffer state of the current slot is fully determined by the channel parameters and arrived packets in the previous slot, as well as the last scheduling decision. This tier can be formulated as a Markov decision process. In addition, as the underlying small-scale dynamics are essentially non-stationary, this tier can be further formulated as a non-stationary Markov decision process, for which the stochastic approximation and reinforcement learning methods can be applied for the model-based and data-based scenarios, respectively.

A Proof of Concept

In this section, we provide a proof of concept of the proposed slicing architecture for RRM in multi-cell multi-slice networks. For the benchmark two-tier architecture, we adopt the RRM schemes proposed in [6]. For the proposed three-tier architecture, we adopt an iterative method to minimize the amount of radio resources suffering from inter-slice interference in the first tier, and adopt the online convex optimization algorithm in [8] to adjust the local radio resource allocation in the middle tier. In the lower tier, the proportional fair scheduler with coordinated multi-point is utilized. In addition, we consider two different versions for each architecture. The high complexity version adopts a shorter execution period than the low complexity version to address more dynamic environments.

As shown in Fig. 4, we show the isolation-capacity-complexity tradeoff for both the proposed three-tier solution and the benchmark two-tier solution. Here, the slice capacity is defined as the ratio of the average number of required resource blocks per cell to the number of available resource blocks. The slice isolation is defined as the percentage of slices that

have a guaranteed number of resource blocks that are free from inter-slice interference. The implementation complexity refers to the high and low versions. As the proposed three-tier architecture allows each tier to focus on network dynamics with specific time and spatial scales, the corresponding RRM methods can achieve higher traffic throughput, better isolation between slices and lower computational complexity. The advantage in isolation-capacity-complexity tradeoff enables to deploy more practical and flexible RRM solutions in future multi-cell multi-slice networks.

Conclusion

In this article, we considered the RRM in multi-cell multi-slice networks. We first show that the problem is highly complicated as the additional slice dimension is tightly coupled with the conventional cell dimension, for which we propose a key design tradeoff among three major aspects, i.e., slice isolation, slice capacity and implementation complexity. Then, we characterize the network dynamics according to their spatial and temporal scales, and propose a hierarchical slicing architecture with three tiers, in which an additional tier is introduced to coordinate the local radio resource allocation between neighboring cells and slices. At last, we provide a proof of concept to show the advantages of the proposed three-tier solution as compared to a benchmark two-tier solution, in terms of the isolation-capacity-complexity tradeoff.

The advantage in isolation-capacity-complexity tradeoff enables to deploy more practical and flexible RRM solutions in future multi-cell multi-slice networks.

References

[1] H. Zhang et al., "Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, Aug. 2017.

[2] T. Wang and S. Wang, "Inter-slice radio resource allocation: An online convex optimization approach," *IEEE Wireless Commun.*, vol. 28, no. 5, pp. 171–177, Oct. 2021.

[3] J. Li et al., "A hierarchical soft RAN slicing framework for differentiated service provisioning," *IEEE Wireless Commun.*, vol. 27, no. 6, pp. 90–97, Dec. 2020.

[4] S. D'Oro, F. Restuccia, and T. Melodia, "Toward operator-to-waveform 5G radio access network slicing," *IEEE Commun. Mag.*, vol. 58, no. 4, pp. 1823, Apr. 2020.

[5] H. Li et al., "An interference minimization-based RAN slicing strategy in 5G systems," in *Proc. 17th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Berlin, Germany, Sep. 2021, pp. 16.

[6] S. D'Oro et al., "Coordinated 5G network slicing: How constructive interference can boost network throughput," *IEEE/ACM Trans. Netw.*, vol. 29, no. 4, pp. 1881–1894, Aug. 2021.

[7] A. Papa et al., "User-based quality of service aware multi-cell radio access network slicing," *IEEE Trans. Netw. Service Manag.*, vol. 19, no. 1, pp. 756–768, Mar. 2022.

[8] T. Wang and S. Wang, "Online convex optimization for efficient and robust inter-slice radio resource management," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6050–6062, Sep. 2021.

[9] N. Nikaein et al., "Network Store: Exploring slicing in future 5G networks," in *Proc. 10th ACM Workshop Mobility Evolving Internet Archit. (ACM MobiArch)*, Paris, France, Sep. 2015, pp. 8–13.

[10] R. Ferrus et al., "On 5G radio access network slicing: Radio interface protocol features and configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, May 2018.

[11] A. Ksentini et al., "Providing low latency guarantees for slicing-ready 5G systems via two-level MAC scheduling," *IEEE Netw.*, vol. 32, no. 6, pp. 116–123, Nov./Dec. 2018.

[12] B. Soret et al., "Interference coordination for 5G new radio," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 131–137, Jun. 2018.

[13] Z. Gu et al., "Knowledge-assisted deep reinforcement learning in 5G scheduler design: From theoretical framework to implementation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2014–2028, Jul. 2021.

[14] C. Song et al., "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[15] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 1–11.

Biographies

Tianyu Wang (Member, IEEE) (wangty@njupt.edu.cn) received the Ph.D. degree from Peking University, Beijing, China, in 2016. From 2017 to 2023, he was with the School of Electronic Science and Engineering, Nanjing University, China. He is currently an Assistant Professor with the School of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China. He received the Best Paper Award from the IEEE ICC'15, IEEE GLOBECOM'14, and ICST ChinaCom'12. His current research interest includes network slicing and machine learning in wireless networks.

Xun Cao (Member, IEEE) (caoxun@nju.edu.cn) received the B.S. degree from Nanjing University, Nanjing, China, in 2006, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2012. He held visiting positions with Philips Research, Aachen, Germany, in 2008, and Microsoft Research Asia, Beijing, from 2009 to 2010. He was a Visiting Scholar with the University of Texas at Austin, Austin, TX, USA, from 2010 to 2011. He is currently a Full Professor with the School of Electronic Science and Engineering, Nanjing University. His research interests include computational photography, image-based modeling and rendering, and VR/AR systems.

Shaowei Wang (Senior Member, IEEE) (wangsw@nju.edu.cn) received the Ph.D. degree from Wuhan University, Wuhan, China, in 2006. He joined the School of Electronic Science and Engineering, Nanjing University, Nanjing, China, as a Faculty Member, in 2006, where he is currently a Full Professor. From 2012 to 2013, he was a Visiting Scholar/Professor with Stanford University, Stanford, CA, USA, and The University of British Columbia, Vancouver, BC, Canada. His research interests include communications and networking, operations research, and machine learning.