A Hierarchical Soft RAN Slicing Framework for Differentiated Service Provisioning

Junling Li, Weisen Shi, Peng Yang, Qiang Ye, Xuemin (Sherman) Shen, Xu Li, and Jaya Rao

Abstract

Network slicing is a key technology to allow resource sharing among heterogeneous operators/services, which achieves QoS isolation for service provisioning in future communication networks. In this article, a comprehensive hierarchical soft-slicing framework is proposed to enable software-defined radio access networks supporting differentiated services with diverse QoS requirements. The proposed framework consists of network-level slicing and gNodeB-level slicing. In the network level, radio RBs are pre-allocated to each gNodeB in a large time scale, while in the gNodeB level, the pre-allocated RBs are dynamically scheduled to the services in response to the small time scale (mini-slot-level) RB request variations. The proposed framework allows the accommodation of time-varying traffic loads of differentiated services over multiple gNodeBs, while enabling dynamic inter-gNodeB RB sharing to increase the resource multiplexing gain. A case study is presented to demonstrate the effectiveness of the proposed framework, followed by a discussion on open research issues.

Introduction

With the prevalence of Internet-of-Things (IoT), a massive number of intelligent devices (e.g., sensors, vehicles, wearable devices) are expected to be connected to accommodate various newly emerging services [1, 2], including ultra-reliable and low-latency communication (URLLC) services, massive machine-type communication (mMTC) services, and enhanced mobile broadband (eMBB) services [3]. These new services demonstrate highly diversified traffic characteristics and differentiated quality-of-service (QoS) requirements. In particular, URLLC is crucial to many prospective applications, such as industry automation, autonomous driving, and remote surgery, which require a low end-to-end (E2E) communication delay in the order of milliseconds. Meanwhile, the reliability requirements of URLLC services can be higher than 99.999 percent [3]. In contrast, eMBB services require high data rates (up to the order of Gb/s) supported on moving devices over a wide coverage area. Typical applications include 4K/8K ultra-high definition video streaming, virtual reality, and augmented reality.

To support the massive number of terminals with seamless connectivity, radio access networks (RAN) are foreseen to consist of a multi-tier of network components, that is, macro-cells, small cells, and femtocells. However, the inherent radio resource scarcity problem poses technical challenges on RAN resource management, considering the increased inter-cell/inter-tier interference. Hence, a fundamental research issue is how to achieve efficient spectrum utilization in a multi-tier RAN, while meeting the diversified service requirements.

As one of the key enabling technologies for future networks, network slicing holds great potential to support various services by logically partitioning the network resources into multiple virtual slices for customized services [4–6]. Typically, each sliced network consists of a RAN slice and a core network slice. Each RAN slice incorporates the radio access and processing functions from a set of base stations (i.e., gNodeBs) and the allocated radio resource blocks (RBs) to support a certain type of service. Each core network slice has a set of network/service-level functionalities (e.g., firewall, transcoding) and associated processing resources and link transmission resources. Multiple network slices coexist over a physical substrate network, and their resources are logically isolated and managed independently by different virtual network operators (VNOs). However, to satisfy the diverse QoS requirements of different slices, the network-wide radio resources have to be partitioned efficiently, in order to maximize the utilities of service providers. Software-defined networking (SDN) is the key technology to enable RAN slicing [1, 3, 5–7], as it realizes programmability on all the gNodeBs, and thus a logically centralized controller is able to manage the radio resources over the whole physical network. With the help of SDN, gNodeB radio resources can be flexibly "sliced" to achieve better QoS provisioning and more efficient resource utilization. However, the signaling overhead between gNodeBs and the controller can be increased if the system information within gNodeBs is frequently updated due to significant traffic fluctuation.

To allow network-wide radio resource sharing among different network slices/VNOs in a cost-effective manner while ensuring stringent and diverse QoS, it is imperative to take advantage of a two-level (i.e., network-level and gNodeB-level) SDN-based radio resource allocation framework to facilitate spectrum exploitation among different network slices in different time scales. With the use of the SDN controller, the network-wide RBs can

Junling Li, Weisen Shi, and Xuemin (Sherman) Shen are with the University of Waterloo; Peng Yang (corresponding author) is with Huazhong University of Science and Technology; Qiang Ye is with Minnesota State University; Xu Li and Jaya Rao are with Huawei Technologies Canada Inc.

Digital Object Identifier: 10.1109/MWC.001.2000010

1536-1284/20/$25.00 © 2020 IEEE

be pre-allocated to each gNodeB in a large time scale, while each gNodeB can dynamically schedule the pre-allocated RBs to each end user in a small time scale, in response to traffic burstiness. By establishing the two time granularity resource allocation policy, the signaling overhead between BSs and devices for the network-level RB re-allocations can be reduced. In this article, we propose a hierarchical "soft" RAN slicing framework that allows the opportunistic radio resource sharing among different network slices over network-wide, while ensuring stringent and diverse QoS guarantee. The proposed framework is designed to accommodate time-varying traffic load by dynamic inter-gNodeB resource sharing to exploit the resource multiplexing gain. The remainder of the article is organized as follows. Motivations and challenges for soft RAN slicing are first discussed. A hierarchical soft-RAN slicing framework is then presented, followed by a case study and a discussion on open research issues. Finally, conclusion remarks are drawn.

Motivation and Challenges

Motivation

The motivation of designing a RAN slicing framework to satisfy differentiated QoS requirements of diverse services in a cost-effective and flexible manner can be elaborated from the following aspects:

QoS Isolation: Different network services are usually with distinct QoS requirements. However, the variation of network states (including user mobility, channel quality, and traffic load) of one service should not affect the QoS provisioning of another service, in order to achieve QoS isolation [6]. To this end, radio resource slicing is necessary to reorganize and partition the whole network resources into slices, each of which is allocated to one service to ensure QoS isolation. To facilitate the spectrum exploitation among different network slices and achieve flexible radio resource management, it is crucial to have a global network view over the available radio resources. SDN is an effective technology to enable centralized control with global network view. By using SDN, the radio resources can be managed flexibly among gNodeBs to realize network-wide resource sharing.

Differentiated Resource Allocation Granularity: Time-varying traffic loads among gNodeBs require dynamic resource allocation to achieve improved resource utilization. However, the signaling overhead between gNodeBs and the SDN controller can become high if traffic states within gNodeBs is frequently updated. To reduce the signaling exchange frequency, it is desired that the RBs can be pre-allocated to each gNodeB in a large time scale to realize coarse-grained service provisioning, while allowing fine-grained RB allocation adjustment in each gNodeB to achieve strict QoS-guarantee. This motivates the design of a hierarchical (i.e., two-level) RAN slicing framework for supporting differentiated services in future networks.

Adaptation with Traffic Variation: To support differentiated services in SDN-enabled RAN, the pre-allocated radio resources to each gNodeB can be inefficient due to the varying traffic load. At the same time, resources among slices are also shared and dynamically updated according to real-time traffic loads. Therefore, some of the gNodeBs can be overloaded while others are under-utilized in terms of RB resources. A soft RAN slicing scheme is required, in which the RBs pre-allocated to one gNodeB need to be temporarily accessed by some of other gNodeBs.

Challenges

Designing a two-level soft RAN slicing framework for supporting diverse services also faces technical challenges:

Strict QoS Guarantee: Some 5G network services such as URLLC services may have strict QoS requirements in terms of delay (less than 1 ms) and reliability (higher than 99.999 percent). To guarantee the ultra-low delay requirements, the radio resources need to be re-scheduled in short time intervals (i.e., within milliseconds or even smaller [8]). To ensure the stringent QoS requirements, the allocated resource should be either over-provisioned, which causes low resource utilization, or frequently updated as often as the time scale of a scheduling interval, which requires significant signaling overhead between gNodeBs in the RAN. Therefore, a RAN slicing scheme needs to balance the trade-off between resource utilization and network dynamic adaption [8, 9], which poses challenges on QoS-aware RAN slicing.

Spatial and Temporal Traffic Dynamics: 5G services usually have differentiated traffic patterns. For instance, traffic arrivals of each URLLC user is highly bursty, while eMBB user data traffic is highly coupled with its mobility pattern and spatial locations [3, 10]. To have an efficient RAN slicing scheme, customized and accurate traffic modeling for each service is a necessity. On the other hand, since the wireless channel conditions and RB scheduling performance are affected by user mobility and traffic dynamics, it is technically challenging to design an appropriate mapping relation between traffic loads and the required resource amounts, given a heterogeneity of users and services.

Signaling Overhead Reduction: Due to high traffic dynamics, it is challenging to achieve optimal network-level resource allocation as frequent interaction between gNodeBs and the SDN controller incurs excessive signaling overhead. On the other hand, resource blocks can be scheduled within each gNodeB at a much smaller time scale (comparable to the duration of one mini-slot) to reduce the information exchange between the controller and gNodeBs. Therefore, a RAN slicing problem should be studied with two spatial-temporal levels considered, where network-wide resource slicing is conducted to achieve global optimality over the whole RAN, and a small time scale resource scheduling is operated to adapt to traffic variations in each gNodeB. How to balance the trade-off between re-slicing signal overhead and global slicing optimality turns out to be an essential issue in designing the two-level RAN slicing scheme.

Multiplexing Gain Exploration: While guaranteeing differentiated service QoS requirements, a non-static resource allocation scheme is expected to improve the multiplexing gain. Existing RAN slicing strategies are either "hard" or "soft." Hard slicing schemes assign a fixed number of RBs to gNodeBs, which ensures QoS isolation without exploring multiplexing gain. In contrast, soft slicing

To facilitate the spectrum exploitation among different network slices and achieve flexible radio resource management, it is crucial to have a global network view over the available radio resources. SDN is an effective technology to enable centralized control with global network view. By using SDN, the radio resources can be managed flexibly among gNodeBs to realize network-wide resource sharing.

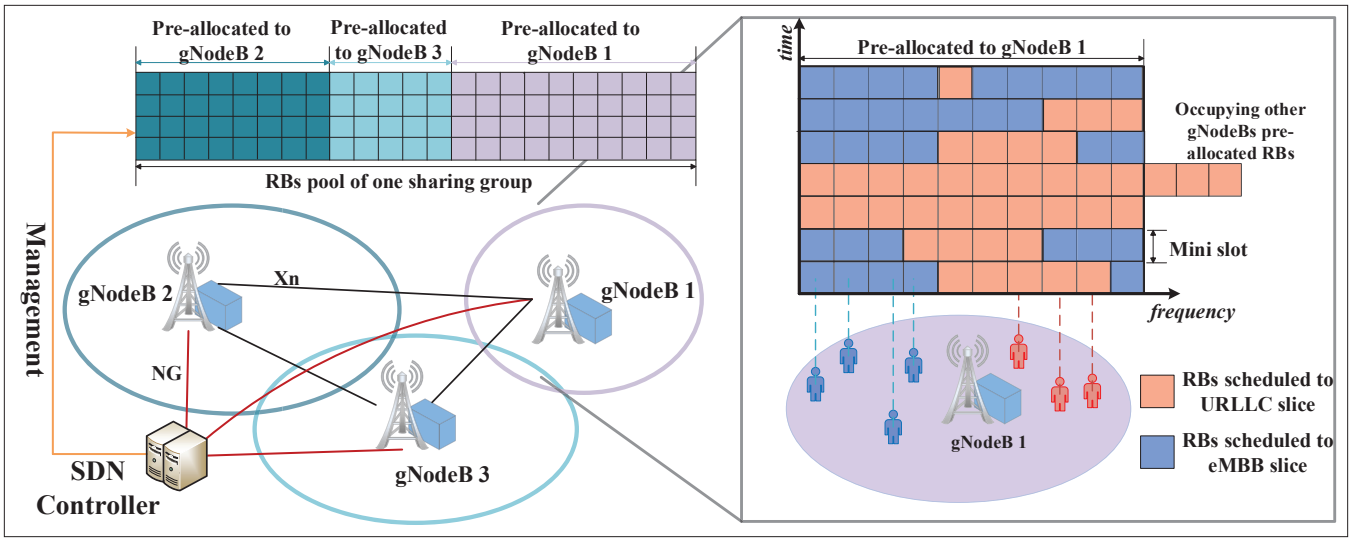

FIGURE 1. Network architecture for the hierarchical soft RAN slicing framework.

schemes enable dynamic sharing of RAN resources among gNodeBs, which realizes flexible resource allocation with high multiplexing. Therefore, a soft radio resource slicing mechanism is effective to balance the trade-off between QoS isolation and multiplexing gain. The main challenge is to determine the optimal ratio of the amounts of reserved resources over shared resources.

Proposed Hierarchical Soft RAN Slicing Framework

Network Architecture

As shown in Fig. 1, we consider two typical types of services for 5G networks, that is, URLLC services and the eMBB services. Four network components exist in the considered network architecture as follows:

SDN Controller: We consider SDN-enabled RAN where a set of gNodeBs in the same network tier are directly connected to and managed by a central SDN controller through the NG interfaces [11]. All the gNodeBs in the RAN share a common radio resource pool and the SDN controller has a centralized control over all radio resources in the gNodeBs.

gNodeB: A gNodeB refers to a 5G base station which accommodates the service requests inside its coverage area. A set of gNodeBs are considered as a gNodeB sharing group if they are highly overlapped in their communication coverage areas. Due to the highly overlapped coverage areas, frequency reuse among gNodeBs in one sharing group is ineffective in reducing inter-gNodeB interference. Therefore, orthogonal resources are allocated to gNodeBs within one sharing group, and frequency reuse takes place among different sharing groups. The gNodeB-level RB scheduling is performed by each gNodeB in a small time scale.

URLLC User: URLLC users generate URLLC service requests that require ultra-high reliability and low latency. Heterogeneous URLLC users with different traffic patterns can be served by one gNodeB. URLLC users are in general with low mobility, so they can be considered as quasi-static within one slicing period. The RBs are scheduled to URLLC services periodically in each URLLC transmit time interval (TTI) with a duration of 0.125 ms [3]. Each URLLC TTI is referred to as a mini-slot.

eMBB User: eMBB users usually generate service requests requiring high average data rates. RBs are scheduled to eMBB services periodically in each eMBB TTI with a duration of 1 ms [8]. Within each eMBB TTI, URLLC traffic has higher priority over eMBB traffic to guarantee these stringent QoS requirements [3]. Therefore, the URLLC traffic is allowed to occupy scheduled RBs for the eMBB traffic, which is also referred to as the URLLC traffic "superposition" over eMBB transmission [8]. Since eMBB services widely exist in both high mobility scenarios (e.g., vehicle/high-speed railway) and in low-to-moderate mobility scenarios (e.g., pedestrian movement), their required RB amount can be different due to mobility patterns, even for two services with the same QoS requirements. Therefore, the mobility features of eMBB users need to be incorporated into the resource slicing problem formulation.

The hierarchical slicing framework consists of two levels, that is, network-level RB pre-allocation among gNodeBs, and gNodeB-level RB scheduling within each gNodeB. During network-level RB pre-allocation, the network-wide RBs are allocated to each gNodeB by the SDN controller in a large time scale. In the gNodeB-level scheduling phase, the pre-allocated RBs are scheduled to the service requests generated by URLLC/eMBB users within each gNodeB in each slot/mini-slot. By separating network-level and gNodeB-level resource allocation, the overhead for frequent RB re-allocations is significantly reduced.

Network-Level Resource Pre-Allocation

The objective of network-level RB pre-allocation is to pre-allocate RBs for all the gNodeBs. Different from the hard slicing schemes that allocate a fixed number of RBs to each gNodeB, the number of RBs assigned to each gNodeB by network-level RB pre-allocation varies in different scheduling slots according to instantaneous traffic demands, which conducts the resource allocation in a "soft" way. This mechanism not only ensures the QoS requirements of differentiated services, but also enables the gNodeBs to share

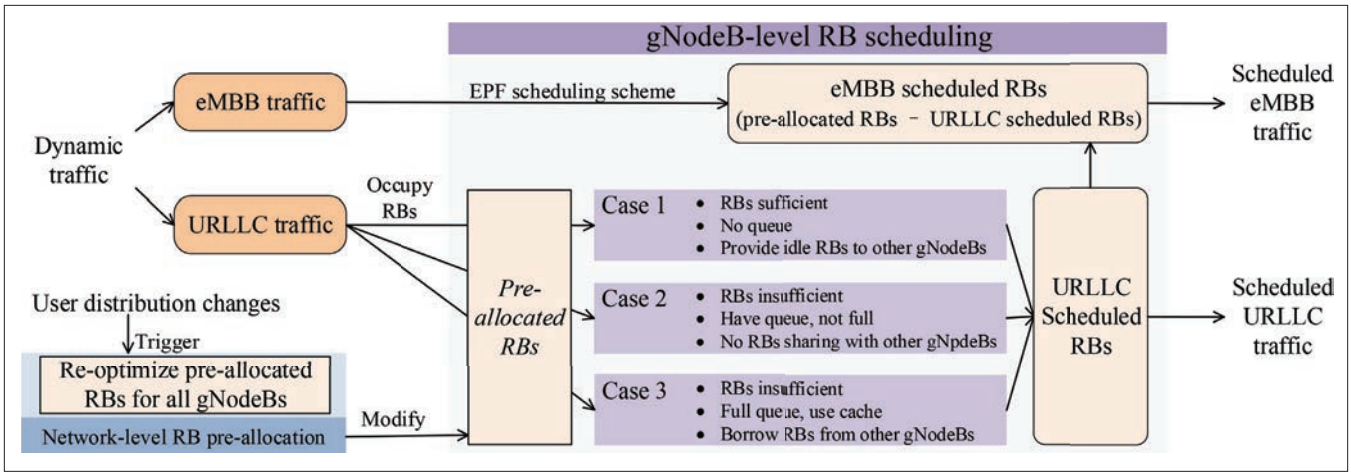

FIGURE 2. An illustration of the proposed gNodeB-level RB scheduling process in each gNodeB.

their idle RBs with other overloaded gNodeBs. On the other hand, compared with the "fully soft" slicing schemes which share the RBs pool among all the gNodeBs, the RBs pre-allocated to each gNodeB can be directly accessed by other gNodeBs without frequent access negotiations or contentions [12, 13]. Those frequent inter-gNodeB negotiations or contend-based access mechanisms are not effective in guaranteeing the delay and reliability QoS requirements in a mini-slot level.

During the network-level resource pre-allocation, the strict delay and reliability requirements of URLLC services, as well as the minimal average throughput of eMBB services, should be satisfied. To this end, the mapping relation between URLLC/eMBB traffic (in packets/mini-slot) and the number of required RBs (in RBs/mini-slot) should be investigated. The aggregate URLLC traffic arrivals at each gNodeB can be modeled as a Markov-Modulated Poisson Process (MMPP) [14]. The delay violation probability for URLLC traffic at each gNodeB is thus analyzed based on queuing theory to ensure the delay and reliability satisfaction. The average RB requirement for an eMBB service is calculated based on the mobility model applied, that is, the random waypoint (RWP) model which can characterize different enhanced mobility patterns based on the model parameter selection [15]. Then, an optimization problem is formulated to minimize the total number of RBs pre-allocated to all gNodeBs in one sharing group. Specifically, the delay requirement for URLLC services can be interpreted as the maximal packet queue length based on packet queuing analysis, and the URLLC reliability requirement constrains the maximal loss probability of URLLC packets during transmission and queuing. The average throughput requirement of eMBB services is expressed as the minimal average number of remaining RBs after URLLC packet scheduling. With those methods, the URLLC and eMBB QoS requirements are incorporated into the problem formulation.

The main advantages for the network-level resource pre-allocation include:

- A "soft-slicing" mechanism is developed which guarantees QoS requirements of differentiated services by non-fixed resources pre-allocated to each gNodeB.
- Both the traffic burstiness and inter-gNodeB resource sharing probability are exploited to improve the network-level multiplexing gain.
- Frequent re-slicing is avoided by allowing gNodeB-level RB scheduling to address dynamic bursts in traffic, which significantly reduces the signaling overheads between the SDN controller and the gNodeBs.

gNodeB-Level RB Scheduling

Given the pre-allocated RBs from network-level RB pre-allocation, each gNodeB executes gNodeB-level RB allocation to schedule the RBs based on the instantaneous traffic rate in each mini-slot. Meanwhile, the RBs requests and sharing between gNodeBs are conducted in a mini-slot level to realize collision-free inter-gNodeB RBs sharing, while satisfying the QoS requirements of differentiated services.

Traditional RB scheduling algorithms, such as the round robin (RR) algorithm and the enhanced proportional fair (EPF) algorithm, can hardly guarantee the strict reliability and latency requirements of URLLC services [12]. Meanwhile, the inter-gNodeB RB sharing, which is essential to ensure the reliability requirement of URLLC services, is not enabled in those algorithms. Therefore, a new gNodeB-level RB scheduling scheme for each gNodeB to assign RBs to users is designed. The new scheme ensures both reliability and latency requirements for URLLC services, and supports dynamic inter-gNodeB and inter-mini-slot RB sharing so as to increase the radio resource multiplexing gain.

The proposed gNodeB-level RB scheduling scheme that enables inter-gNodeB RB sharing is illustrated in Fig. 2. The gNodeB-level RB scheduling scheme is executed on each gNodeB. Each gNodeB maintains a transmission queue. For eMBB traffic, the traditional EPF scheme is applied to ensure the average throughput requirements and the fairness between services. Note that URLLC data can occupy the scheduled RBs to eMBB services within each eMBB TTI. For multiple eMBB services with the same priority, URLLC data traffic randomly occupy the RBs scheduled for eMBB traffic; for eMBB services with different priorities, the scheduled RBs are occupied by URLLC data traffic with different probabilities for QoS prioritization.

Traditional RB scheduling algorithms can hardly guarantee the strict reliability and latency requirements of URLLC services. Meanwhile, the inter-gNodeB RB sharing, which is essential to ensure the reliability requirement of URLLC services, is not enabled in those algorithms. Therefore, a new gNodeB-level RB scheduling scheme for each gNodeB to assign RBs to users is designed.

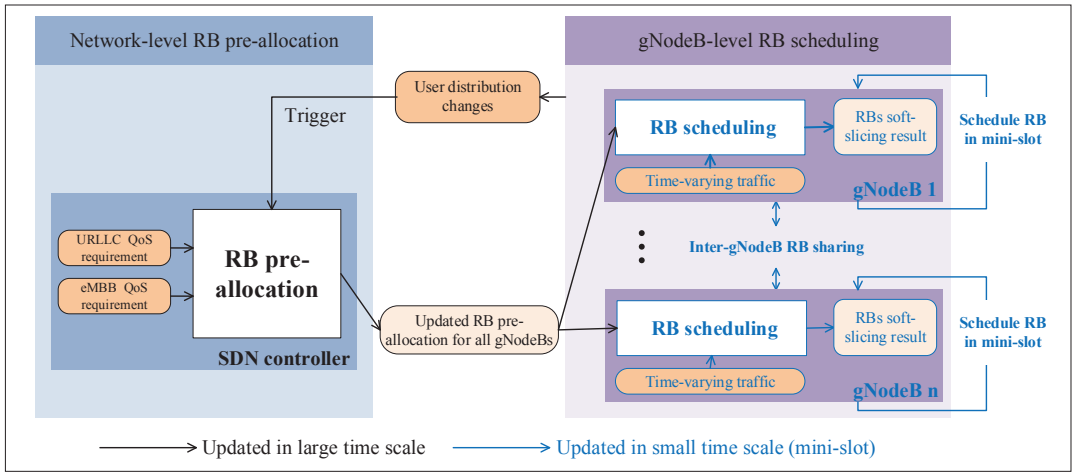

FIGURE 3. An illustration of the proposed hierarchical soft RAN slicing framework.

To guarantee the delay tolerant constraint of URLLC services, the queue length threshold is designed to be the number of pre-allocated RBs in each mini-slot times the maximal allowed latency for URLLC service in unit of mini-slot. At the beginning of each mini-slot, each gNodeB determines the scheduling decision according to the traffic loads.

Case 1: The pre-allocated RBs in one mini-slot are sufficient to support newly arrived and existing URLLC data in gNodeB's transmission queue (in unit of RB), all the required RBs are scheduled and then being transmitted in the next mini-slot. The eMBB data are scheduled for transmission using remaining RBs allocated for URLLC traffic, and the traditional EPF algorithm is applied.

Case 2: When the number of newly arrived and queued data (in unit of RB) is larger than the pre-allocated RBs amount in one mini-slot, but smaller than the maximal queue length, the gNodeB schedules all the pre-allocated RBs for transmitting data in the queue on a first-in-first-out basis. The remaining URLLC data are added to the end of the queue.

Case 3: The maximal queue length is violated, the gNodeB will cache the extra data, and broadcast a message requesting the same number of RBs from other gNodeBs in the sharing group. Through the interaction among gNodeBs in one sharing group, the proposed gNodeB-level RB scheduling scheme can realize collision-free inter-gNodeB RBs sharing, so as to increase the resource multiplexing gain.

The gNodeB-level RB scheduling has the following advantages:

- Realize a collision-free inter-gNodeB negotiation mechanism which is available for mini-slot level RBs scheduling for URLLC services.
- Addressing traffic bursts locally instead of informing the controller for re-slicing.
- Cooperate with network-level resource pre-allocation to ensure QoS requirements of differentiated services in real-time operation.

Operation Workflow

Figure 3 shows the interaction between the two-level resource slicing in the proposed framework. The network-level and gNodeB-level mechanisms are implemented on the SDN controller and each gNodeB, respectively, and are running in different time granularities.

First, the SDN controller performs the network-level RB pre-allocation based on the service requests of all gNodeBs in one sharing group. After the pre-allocation phase is completed, each gNodeB monitors the user distribution inside its coverage area. For the variation of traffic statistics (e.g., user distribution), the gNodeBs report new traffic states to the controller and a re-execution of network-level RB pre-allocation can be triggered. Afterwards, the updated decision of RB pre-allocation is sent to all the gNodeBs. For temporal traffic bursts that can be addressed by gNodeB-level RB scheduling (e.g., URLLC traffic bursts, eMBB user moving), the gNodeBs avoids to refer to the SDN controller for the network-level resource reallocation. Then, for the gNodeB level resource scheduling, each gNodeB makes the scheduling decision in mini-slot, and exchanges RBs with other gNodeBs through inter-gNodeB negotiation. The number of total RBs to be scheduled within one mini-slot for each gNodeB is updated after each re-execution of network-level RB pre-allocation.

The proposed framework has the following main advantages:

- Fast response in the gNodeB-level resource scheduling and global optimality in the network-level resource allocation, with reduced controller-gNodeB communication overhead.
- Guaranteed QoS requirements of differentiated services and improved resource utilization through dynamic inter-gNodeB RBs sharing.
- Collision-free gNodeB-level RB scheduling via inter-gNodeB negotiation with stringent QoS guarantee for URLLC services.

Case Study

In this section, a case study is presented to evaluate the performance of the proposed soft RAN slicing framework.

Simulation Setting

We consider an SDN-enabled RAN with multiple gNodeBs deployed as a sharing group. Two scenarios are evaluated, that is, Scenario 1 with three heterogeneous gNodeBs and Scenario 2 with three (or five) homogeneous gNodeBs. Two types of mobility patterns (vehicle mobility and pedestrian mobility) are considered for one eMBB service. In Scenario 1, three heterogeneous

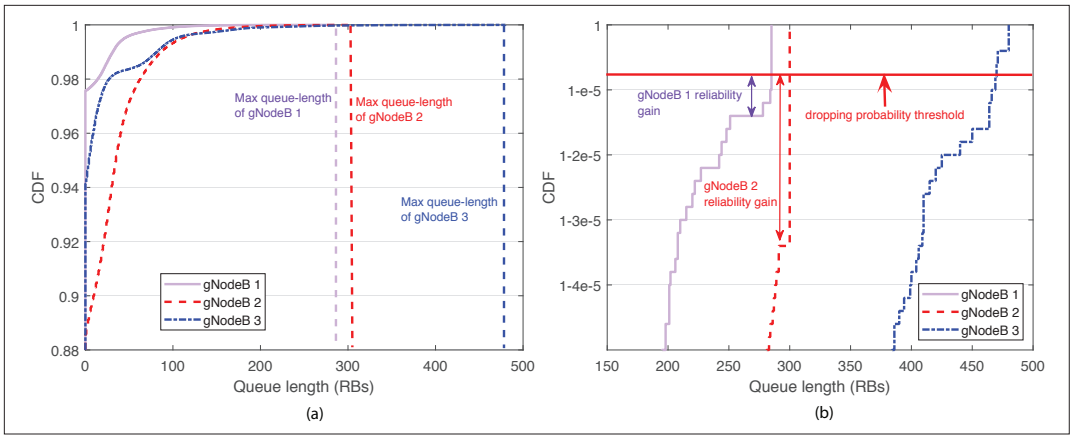

FIGURE 4. The CDF of queue length at the three heterogeneous gNodeBs.

gNodeBs have different numbers of URLLC users and network traffic loads. In Scenario 2, the number of URLLC users and traffic statistics for all the gNodeBs are the same. The whole simulation time is 5 x 105 mini-slots. Other simulation parameters are summarized in Table 1.

Simulation Results

The proposed framework is first evaluated in terms of ensuring the strict delay and reliability requirement of URLLC services in Scenario 1. Figure 4a shows the cumulative distribution functions (CDFs) of the queue length at the three heterogeneous gNodeBs. Since the number of pre-allocated RBs can be interpreted as the service rate per mini-slot, the maximal queue length indicates the bound for URLLC packet delay guarantee. In Fig. 4b, we observe for both gNodeB 1 and gNodeB 2, their packet dropping probabilities are greater than the URLLC reliability threshold. To guarantee the packet dropping probability bound, more RBs are needed from other gNodeBs through inter-gNodeB RB-sharing, which contributes to the reliability gains shown in Fig. 4b.

Next, the homogeneous scenario is considered to compare the proposed soft slicing framework with the hard slicing scheme. The "average RBs for eMBB" curves in all sub-figures of Fig. 5a represent the eMBB performance where the least allowable number of RBs is pre-allocated to one gNodeB, that is, the 5-gNodeB soft-slicing case. Figures 5a-5c show the performance of the hard-slicing scheme and the proposed soft-slicing scheme with different number of gNodeBs. The average number of RBs per gNodeB obtained using the hard-slicing scheme is higher than or equal to that obtained using the proposed soft-slicing scheme. This is because the proposed scheme allows the radio resources to be shared among gNodeBs, leading to a higher resource utilization. In particular, the comparison between the two slicing schemes with the variation of the maximal allowed dropping probability is shown in Fig. 5a. It is seen that the hard-slicing scheme consistently requires more (about 10 percent) RBs than the proposed soft-slicing scheme for all the cases. Moreover, the average number of RBs per gNodeB increases as the dropping probability constraint becomes tighter for both schemes. The gap between the two schemes enlarges as the dropping probability becomes stricter, which demonstrates the advantage of the proposed soft-slicing scheme in terms of ensuring the strict reliability constraints of URLLC services. Figure 5b shows that the average number of RBs per gNodeB increases with the number of URLLC users in each gNodeB for both schemes, while the proposed scheme always saves less RBs per gNodeB than the hard-slicing scheme in all the scenarios. Figure 5c shows the average number of RBs per gNodeB decreases as the delay tolerant constraint becomes loose. This is reasonable as a tighter delay constraint typically demands a larger number of RBs to achieve a shorter queuing delay. The average RBs scheduled for eMBB services shown in Figs. 5a–5c are always higher than the minimal requirement, indicating the effectiveness of the proposed soft-slicing scheme in guaranteeing the eMBB QoS.

Parameters	Values
gNodeB transmit power P_t	33 dBm
Noise spectral density N_0	–174 dBm/Hz
URLLC scheduling interval t_{urllc}	0.125 ms
URLLC maximal allowed delay	0.375 ms
RB bandwidth B	1.44 MHz
Packet size	256 bits
Service caching space C_g^s	60 RBs
Minimal average throughput for eMBB user R_{embb}	30 Mb/s
Average number of eMBB users per gNodeB g (high-speed, low-speed)	(2, 6)
Vehicle-mobility speed range	[40, 60] km/h
Pedestrian-mobility speed range	[1, 4] m/s
Maximal pause time (high-speed, low-speed)	(1, 5) minutes
gNodeB coverage radius R_g	100 m
Number of interference sources per gNodeB	3
User-to-interfering gNodeB distance	U (250, 500) m[1]

[1] $U(a, b)$ means uniformly distributed between a and b.

TABLE 1. Simulation parameters.

The proposed framework has the following main advantages: fast response in the gNodeB-level resource scheduling and global optimality in the network-level resource allocation, with reduced controller-gNodeB communication overhead; guaranteed QoS requirements of differentiated services and improved resource utilization through dynamic inter-gNodeB RBs sharing; and collision-free gNodeB-level RB scheduling via inter-gNodeB negotiation with stringent QoS guarantee for URLLC services.

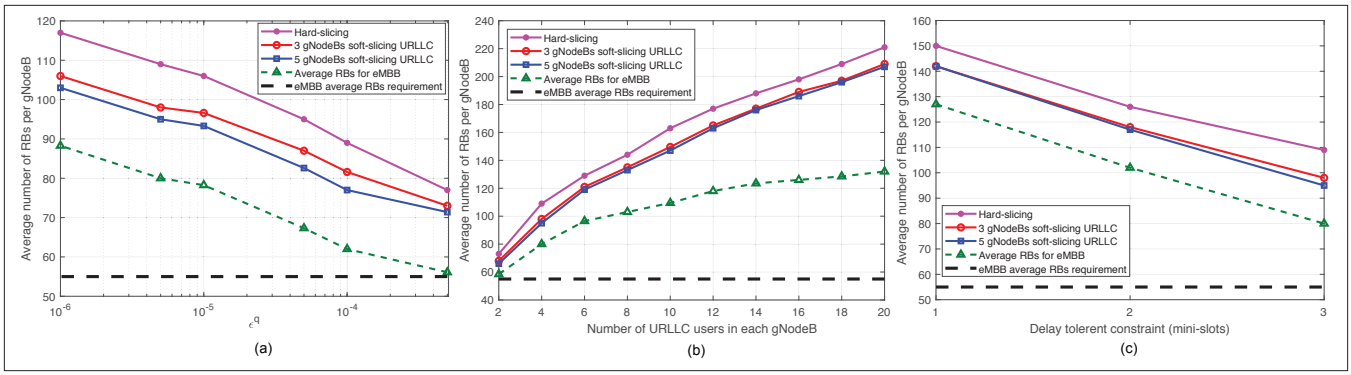

FIGURE 5. Performance comparison: a) average number of RBs per gNodeB as the maximal allowed dropping probability varies; b) average number of RBs per gNodeB vs. number of URLLC users in each gNodeB; c) average number of RBs per gNodeB as delay tolerant constraint of the URLLC traffic changes.

Open Research Issues

In this section, we discuss some open research issues related to soft RAN slicing for future networks.

Prediction-Based Proactive RAN Slicing

The pre-allocated RB amount for both URLLC and eMBB services are highly related to their traffic and mobility patterns. In the proposed scheme, the widely adopted traffic and mobility models (i.e., MMPP for URLLC traffic modeling, RWP for eMBB user mobility) are used for the network-level RB pre-allocation, given the assumption that the user traffic and mobility patterns are well characterized by the models. However, the uncertainty of traffic features and mobility patterns becomes more dominant in an increasingly complex 5G and beyond 5G RAN environment, which decreases the accuracy of model based methods. For future works, a learning based soft RAN slicing solution can be designed to determine the resource allocation. Leveraging the emerging deep learning technologies, such as long short term memory neural network (LSTM) and deep Q-network (DQN), dynamic traffic and mobility patterns of different users or services are trained and predicted accurately with low complexity, upon which proactive resource allocation actions can be learned in an online manner.

Cost-Efficient RAN Slicing

To fully unleash the potential of the hierarchical RAN slicing framework, it is imperative to design a multiple time-granularity resource slicing scheme to balance the network performance and the slicing cost. In particular, both the communication frequency between the SDN controller and gNodeBs, and the information exchange frequency among gNodeBs, have significant impact on the performance of the proposed slicing framework. Moreover, the amount of information required to be exchanged among different network entities for updating the slicing results also affect the slicing accuracy. Higher interaction frequency between different network elements leads to better adaptiveness, but inevitably increases the slicing cost. Therefore, it is required to develop a multi-time-granularity slicing solution to balance the trade-off between slicing cost and slicing accuracy.

Supporting Multi-Connectivity and Incremental Deployment

As a potential solution to improve radio resource utilization, the multi-connectivity requirements need to be incorporated in the proposed hierarchical framework. With multi-connectivity, the user plane data transmission latency of URLLC services can be further decreased since multiple wireless links can be used simultaneously to transmit data. Enabling multi-connectivity further exploits resource utilization by supporting one user with multiple gNodeBs' available resources. However, how to conduct RAN slicing among gNodeBs associated by leveraging BS-user multi-connectivity to guarantee the user's QoS requirements calls for further investigations. The inter-gNodeB synchronization scheme has to be designed to orchestrate resources from different gNodeBs.

As a temporary solution to 5G standalone (SA), 5G non-standalone (NSA) aims to realize the spectrum utilization and physical implementation on currently deployed 4G base stations and communication infrastructure. Considering the coexistence of 4G/5G/B5G RAN in the long run, the RAN slicing framework should be backward compatible in different network scenarios involving heterogeneous radio access technologies. To this end, differentiated services can be grouped and associated with different wireless access technologies/mechanisms based on service characteristics and QoS requirements. For example, 5G-enabled devices can utilize 5G technologies for high data rate services (e.g., eMBB) while still preserving the coexistence of 4G services. Radio resources are expected to be pooled and shared to enable RAN slicing among heterogeneous access technologies. How to conduct resource slicing under a unified communication infrastructure to promote the 5G non-standalone paradigm, considering the difference in resource type, service type, and device type for 4G and 5G, needs deeper investigation.

Conclusion

In this article, we have proposed a comprehensive soft slicing framework to support time-varying traffic loads and enable resource sharing among differentiated gNobeBs. The network-level RB pre-allocation over multiple gNobeBs guarantees QoS requirements with sharing probabilities considerations, then the gNodeB-level RB scheduling

on each gNodeB realizes dynamic scheduling and inter-gNodeB RB sharing. A case study has been presented to demonstrate the effectiveness of the proposed framework in terms of efficient resource utilization.

References

[1] I. Parvez *et al.*, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 4, May 2018, pp. 3098–3130.

[2] L. Wang *et al.*, "Edge-Assisted Stream Scheduling Scheme for the Green-Communication-Based IoT," *IEEE Internet Things J.*, vol. 6, no. 4, 2019, pp. 7282–92.

[3] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-Reliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proc. IEEE*, vol. 106, no. 10, Oct. 2018, pp. 1834–53.

[4] Z. Xiong *et al.*, "Deep Reinforcement Learning for Mobile 5G and Beyond: Fundamentals, Applications and Challenges," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, June 2019, pp. 44–52.

[5] Q. Ye *et al.*, "Dynamic Radio Resource Slicing for a Two-Tier Heterogeneous Wireless Network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, Oct. 2018, pp. 9896–9910.

[6] I. Afolabi *et al.*, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 3, Third Qtr. 2018, pp. 2429–53.

[7] S. Zhang *et al.*, "Air-Ground Integrated Vehicular Network Slicing with Content Pushing and Caching," *IEEE JSAC*, vol. 36, no. 9, Oct. 2018, pp. 2114–27.

[8] A. Anand, G. de Veciana, and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," *Proc. IEEE INFOCOM'18*, Apr. 2018, pp. 1970–78.

[9] O. Sallent *et al.*, "On Radio Access Network Slicing From a Radio Resource Management Perspective," *IEEE Wireless Commun.*, vol. 24, no. 5, Oct. 2017, pp. 166–74.

[10] Z. Hou *et al.*, "Burstiness-Aware Bandwidth Reservation for Ultra-Reliable and Low-Latency Communications in Tactile Internet," *IEEE JSAC*, vol. 36, no. 11, Nov. 2018, pp. 2401–10.

[11] 3GPP TR 38.801 V14.0.0, "Study on New Radio Access Technology: Radio Access Architecture and Interfaces," 2017.

[12] T. Guo and A. Suárez, "Enabling 5G RAN Slicing with EDF Slice Scheduling," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, Mar. 2019, pp. 2865–77.

[13] M. Condoluci *et al.*, "Soft Resource Reservation for Low-Delayed Teleoperation over Mobile Networks," *IEEE Access*, vol. 5, May 2017, pp. 10445–55.

[14] W. Fischer and K. Meier-Hellstern, "The Markov-Modulated Poisson Process (MMPP) Cookbook," *Performance Evaluation*, vol. 18, no. 2, Sept. 1993, pp. 149–71.

[15] M. A. Al Masri and A. B. Sesay, "Mobility-Aware Performance Evaluation of Heterogeneous Wireless Networks with Traffic Offloading," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, Oct. 2016, pp. 8371–87.

Biographies

Junling Li [S'18] received the B.S. degree from Tianjin University, Tianjin, China, in 2013, and the M.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. Her current research interests include SDN, NFV, network slicing for 5G networks, machine learning, and vehicular networks. She received the Best Paper Award at the IEEE/CIC Int'l. Conf. Communications in China (ICCC) in 2019.

Weisen Shi [S'15] received the B.S. degree from Tianjin University, Tianjin, China, in 2013, and the M.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016. He is pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include drone communication and networking, NFV, and vehicular networks. He received the Best Paper Award at the IEEE/CIC Int'l. Conf. Communications in China (ICCC) in 2019.

Peng Yang [S'16, M'18] received his B.E. degree in communication engineering and the Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2013 and 2018, respectively. He was with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, as a visiting Ph.D. student from September 2015 to September 2017, and a postdoctoral fellow from September 2018 to December 2019. Since January 2020, he has been a faculty member with the School of Electronic Information and Communications, HUST. His current research focuses on software defined networking, mobile edge computing, video streaming and analytics.

Qiang Ye [S'16, M'17] has been an assistant professor with the Department of Electrical and Computer Engineering and Technology, Minnesota State University, Mankato, MN, USA, since September 2019. He received his Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2016. He had been with the Department of Electrical and Computer Engineering, University of Waterloo, as a post-doctoral fellow and then a research associate from December 2016 to September 2019. His current research interests include 5G networks, SDN/NFV, network slicing, AI and machine learning for future networking.

Xuemin (Sherman) Shen [M'97, SM'02, F'09] is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, social networks, 5G and beyond, and vehicular ad hoc networks. He is a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Chinese Academy of Engineering Foreign Fellow. He received the R.A. Fessenden Award in 2019 from IEEE, Canada; the James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society; and the Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society.

Xu Li is a senior principal researcher at Huawei Technologies Canada. He received a Ph.D. (2008) degree from Carleton University, an M.Sc. (2005) degree from the University of Ottawa, and a B.Sc. (1998) degree from Jilin University, China, all in computer science. Prior to joining Huawei, he worked as a research scientist (with tenure) at Inria, France. His current research interests are focused in 5G. He has contributed extensively to the development of 3GPP 5G standards through 90+ standard proposals. He has published 100+ refereed scientific papers and is holding 30+ issued U.S. patents. He is/was on the editorial boards of *IEEE Communications Magazine*, *IEEE Transactions on Parallel and Distributed Systems*, *Wiley Transactions on Emerging Telecommunications Technologies* and a number of other international archive journals.

Jaya Rao received his B.S. and M.S. degrees in electrical engineering from the University of Buffalo, New York, in 2001 and 2004, respectively, and his Ph.D. degree from the University of Calgary, Canada, in 2014. He is currently a senior research engineer at Huawei Technologies Canada, Ottawa. Since joining Huawei in 2014, he has worked on research and design of CIoT, URLLC and V2X based solutions in 5G New Radio. He has contributed for Huawei at 3GPP RAN WG2, RAN WG3, and SA2 meetings on topics related to URLLC, network slicing, mobility management, and session management. From 2004 to 2010, he was a research engineer at Motorola Inc. He was a recipient of the Best Paper Award at IEEE WCNC 2014.

Radio resources are expected to be pooled and shared to enable RAN slicing among heterogeneous access technologies. How to conduct resource slicing under a unified communication infrastructure to promote the 5G non-standalone paradigm, considering the difference in resource type, service type, and device type for 4G and 5G, needs deeper investigation.