Dynamic Network Slicing Control Framework in AI-Native Hierarchical Open-RAN Architecture

Jongwon Han, Minhyun Kim†, Jung Mo Moon† and Jeongho Kwak
DGIST, Department of Electrical Engineering and Computer Science, Daegu, South Korea
†*ETRI, Intelligent Small Cell Research Section, Daejeon, South Korea*
{herowhddnjs, jeongho.kwak}@dgist.ac.kr, {minhyun.kim, jmmoon}@etri.re.kr

***Abstract*—Network slicing is a promising technology in next-generation wireless networks that enables the division of a physical network infrastructure into multiple virtual networks, each of which is tailored for specific service requirements. This approach enables a more flexible allocation of network resources such as beamforming vector, bandwidth and transmit power; thereby effectively supporting services that require high data transmission rates. However, in dynamic network environments where multiple users dynamically move around; hence the interference relationships are dynamically varying, traditional static network slicing solution has critical drawbacks. To this end, for the effective implementation and performance improvement in practical and dynamic network environments, we first propose a dynamic network slicing control framework in AI-native hierarchical Open-RAN (Radio Access Network) architecture where mobility prediction and network controls are designed by multiple timescale decomposition. The proposed framework can facilitate effective network controls, enabling the generation of finely tuned QoS management decisions (power/ bandwidth allocation, user scheduling, beam activation) in different timescales. On top of this framework, we compare the performance of a simple dynamic network slicing algorithm and an existing static network slicing scheme via simulations.**

***Index Terms*—Open-RAN, network slicing, resource allocation, GoB beamforming, interference management**

I. Introduction

In recent years, the widespread adoption of smartphones, tablets, and IoT devices, along with the growing popularity of applications requiring substantial data transmission such as Ultra-High Definition (UHD) video streaming, Virtual Reality (VR), and Augmented Reality (AR), has not only lead to a significant increase in mobile data traffic but also increase the demand for higher data transmission rates among user equipments (UEs) [1]. These surges present a significant challenge for current network infrastructures to support.

To cope with these challenges, network slicing has emerged as a promising technology capable of fulfilling the Quality of Service (QoS) requirements of users that demand high data transmission rates. Fully supported by the 5G Open-Radio Access Network (O-RAN) architecture, network slicing enables the virtual segmentation of physical network into multiple, distinct "slices". Each slice can be individually tailored to meet the diverse needs of different users, ensuring more flexible resource allocation and enhanced overall network performance [2].

However, in dynamic network environments where multiple users keep moving, thereby leading to dynamically varying interference relationships, traditional network slicing approaches have critical drawbacks. Therefore, to achieve efficient allocation of network resources and to effectively improve overall performance in such environments, it is crucial to dynamically control the network variables such as bandwidth, transmit power or user scheduling in response to potentially time-varying inter-cell and intra-cell interferences.

Massive MIMO (M-MIMO) has emerged as a promising approach to mitigate interference, employing directional beams to concentrate signals on specific targets, which in turn reduces signal spread loss and minimizes interference [3]. There are generally two types of beamforming methods in M-MIMO systems: Adaptive beamforming and Grid of Beams (GoB) beamforming [4]. Adaptive beamforming utilizes precise channel estimations from Sounding Reference Signals (SRS) sent by users, enabling the achievement of higher data transmission rates. However, *(i)* its implementation is complex and challenging, and *(ii)* it is not well-suited for high-speed users, as the channel estimations become quickly outdated due to the rapid changes in the channel conditions associated with user mobility. On the other hand, the GoB beamforming utilizes a predefined grid pattern of beams for signal transmission, which simplifies the implementation and management of beamforming. Moreover, the GoB beamforming can efficiently transmit signals without the need for complex channel estimations by leveraging the Reference Signal Received Power (RSRP) report from users. This method is especially effective for users moving with high-speed, as it maintains signal consistency regardless of user mobility. Therefore, the GoB beamforming is highlighted as a viable solution for the implementation of M-MIMO beamforming.

However, the joint optimization of GoB beamforming and dynamic network slicing—which involves a dynamic determination of the resource allocation, i.e., transmit power and bandwidth, along with beam activation and user scheduling—require considerable computational complexity [5] [6] [7]. Therefore, to effectively integrate network slicing into GoB beamforming, it is crucial to efficiently determine these network control decisions through low-complex algorithms, and to be supported by an intelligent system that can effectively improve the performance of these algorithms.

In this paper, we propose a dynamic network slicing control framework in AI-native hierarchical O-RAN architecture where mobility prediction and network controls are co-

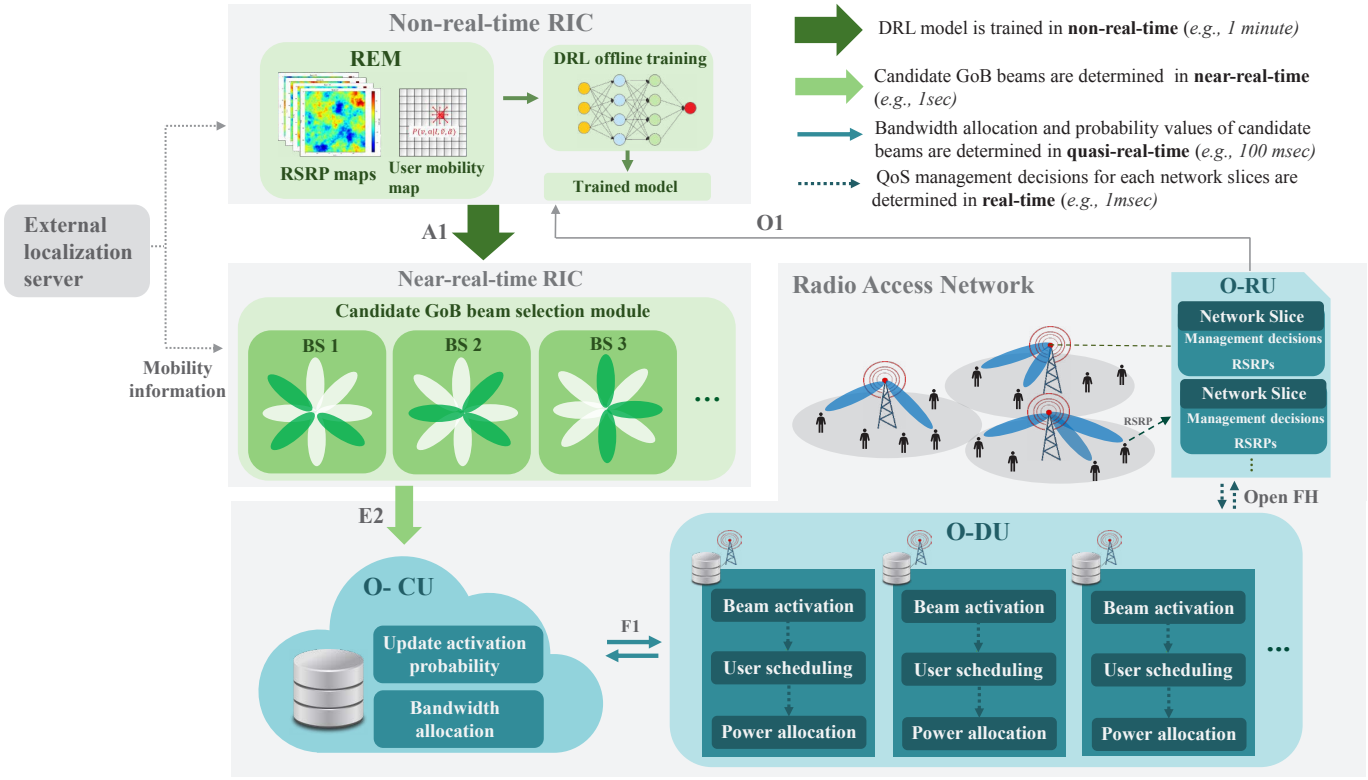

Fig. 1. Proposed dynamic network slicing framework

designed by multiple timescale decomposition. The proposed framework can facilitate efficient decision-making for GoB beamforming and bandwidth, user scheduling and transmit power in divided time slots. Finally, we evaluate a simple dynamic network slicing algorithm on top of the proposed framework compared to the existing static network slicing scheme via simulations. The contributions of this paper are summarized as follows.

- We propose a dynamic network slicing control framework that facilitates efficient decision-making tailored to the given AI-native hierarchical Open-RAN architecture with multiple timescales.
- We evaluate a simple and dynamic network slicing algorithm on top of the proposed framework compared with the existing static network slicing scheme via simulations.

In the rest of this paper, we begin with the description of a dynamic network slicing control framework in Section II. Then, we provide performance evaluation of the proposed framework in conjunction with dynamic network slicing algorithm in Section III. Finally, we conclude this paper in Section IV.

II. Dynamic Network Slicing Control Framework

A. Background Knowledge of O-RAN

O-RAN is an architecture that disaggregates the traditional next-generation NodeB (gNB) into various functional components and connects them using open interfaces [8]. In O-RAN, the conventional RAN is divided into the Radio Unit (RU), Distributed Unit (DU), and Central Unit (CU). The RU primarily handles tasks related to the lower Physical (PHY) layer, such as precoding and beamforming, while the DU manages tasks associated with the upper PHY, Medium Access Control (MAC), and Radio Link Control (RLC) layers, including scrambling and modulation. The CU is responsible for tasks related to the higher layers, such as Radio Resource Control (RRC), Packet Data Convergence Protocol (PDCP), and Service Data Adaptation Protocol (SDAP) layers.

In O-RAN, RAN components are controlled in a software-defined manner using RAN Intelligent Controllers (RICs). RICs are programmable components capable of executing closed-loop control and coordinating the RAN. There are two types of RICs: Near-Real-Time RIC (Near-RT RIC) and Non-Real-Time RIC (Non-RT RIC). Non-RT RIC handles tasks that are relatively more time-consuming, orchestrating the training of AI/ML models for policy management and resource optimization planning. This is essential for facilitating long-term network optimization and strategic decision-making. In contrast, Near-RT RIC supports RAN functions that demand quick responses, using AI/ML algorithms to make immediate decisions of dynamic resource management, interference management, and QoS control.

B. Dynamic Network Slicing Framework

Fig. 1 illustrates a proposed dynamic network slicing framework, which is designed to effectively reduce the high computational complexity involved in dynamically determining

resource allocation of transmit power and bandwidth, along with beam activation, and user scheduling. The proposed framework leverages the hierarchical structure of O-RAN, where each layer operates according to its specific timescale and exchanges information with other layers through the O-RAN interfaces, thereby enabling more efficient network control. The operations at each layer are as follows.

1) Non-RT-RIC: Non-RT RIC is tasked with the training of Deep Reinforcement Learning (DRL) model in Non-real-time. This model is trained to predict the locations of users based on their current locations, and to determine the corresponding set of beams to be activated accordingly. Non-RT RIC collects user mobility information—including location, speed, and direction-from an external localization server [9]. This mobility information, along with the RSRP values for each beam obtained from every users, is used to update the Radio Environment Map (REM). REM is a specialized database that provides two distinct map representations: one for Reference Signal Received Power (RSRP) and another for user mobility patterns [4]. The user mobility map probabilistically represents the movement trends of the users, while the RSRP map displays the RSRP values for each beam across all regions, derived from channel estimation techniques based on user-reported RSRP values. Then, the updated REM is utilized to train a DRL model in order to determine the set of beams to be activated according to the objectives of network operators. For example, it selects beams with the highest RSRP values or chooses beams that balance RSRP values with potential interference. Once training is complete, the trained model is stored in the database of Non-RT-RIC and it is transmitted to the candidate GoB beam selection module in Near-RT RIC via A1 interface every non-real-time period.

2) Near-RT-RIC: In the candidate GoB beam selection module within Near-RT-RIC, an output of the trained DRL model selects candidate beams by analyzing the current locations of users, as provided by an external localization server. The term "candidate beams" refers to the beams chosen as eligible for activation, with activation decision confined to these beams exclusively. The information of the candidate beams determined for each network slice are transmitted to the O-CU via E2 interface every near-real-time period.

3) O-CU: O-CU deployed at the Multi-access Edge Computing (MEC) server assigns activation probabilities to each candidate beam received from Near-RT RIC. These probabilities are determined according to the dynamic network slicing algorithm that is integrated within the proposed framework, and are periodically updated for performance improvement.

Bandwidth allocation is also determined by the O-CU. It is fine-tuned for each network slice by leveraging multiple network control decisions determined in O-DU. By placing the decision-making for bandwidth allocation at a higher layer with longer update periods than other management decisions, i.e., beam activation, user scheduling and power allocation, this architecture can significantly reduce the computational complexity while providing more macroscopic perspective in decision-making of the bandwidth allocation. Then, the determined beam activation probabilities and bandwidth allocation are transmitted to the O-DU in every quasi-real-time period.

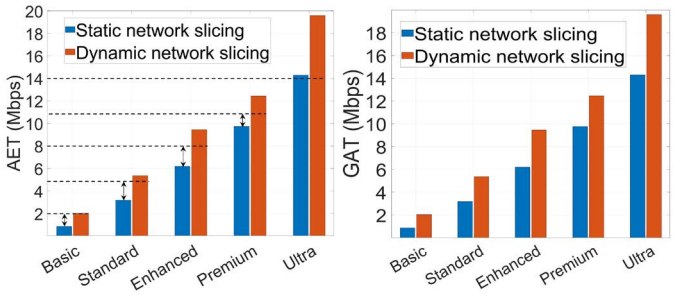

Fig. 2. AET and GAT for each network slice.

4) O-DU: In the proposed framework, each gNB has a dedicated edge computing server [10]. In O-DU deployed in this server, beam activation, user scheduling, and power allocation are made sequentially for each network slice in real-time. The beam activation is determined probabilistically from among the candidate beams, based on their probabilities assigned by the O-CU. For given activated beams, user scheduling and power allocation are determined to meet the requirements of each network slice. We should be noted that all decision variables can be obtained according to the corresponding algorithms in the proposed framework. These algorithms can be developed using optimization and/or learning theory or heuristic methods. In this paper, proposal of a specific algorithm is out of scope, but we provide a novel dynamic network slicing framework. Nevertheless, we exploit a simple dynamic network control algorithm in simulations to show the superiority of the proposed framework. Then, a tuple of the determined control variables is delivered to O-RU via open fronthaul link.

5) O-RU: O-RU is deployed at the cell site and utilizes network control decisions received from O-DU to transmit signal to users. Additionally, each O-RU transmits RSRP reports of users corresponding to reference signal of each beam to Non-real-time RIC via O1 interface.

III. Performance Evaluation

We evaluate the performance of the proposed framework in conjunction with simple dynamic/static network slicing algorithms. We provide the simulation setup and brief description of two network slicing algorithms. As performance metrics, we consider AET and GAT where AET is the average throughput of users whose throughputs are bottom 5% of all users and GAT is the geometric average throughput which captures both throughput and fairness among users [11].

A. Simulation Setup

In our simulations, the network topology includes 19 gNBs, and each gNB is partitioned into 5 distinct network slices according to QoS requirement levels. The network slices are categorized into various priority levels: 'Basic', 'Standard', 'Enhanced', 'Premium', and 'Ultra', each of which corresponds to [2, 5, 8, 11, 14] Mbps of average data rate, and

each network slice randomly serves 4 to 8 users. Users are initially placed randomly and subsequently choose to either remain stationary (0 m/s), walk (1.3 m/s), or run (2.5 m/s) in a random direction. All gNBs are configured with a total bandwidth of 100MHz, and the maximum power is set to 2W. The simulation is executed for 5000 time slots where time intervals of near-real-time, quasi-real-time, real-time are set to 1000, 100 and 1 time slots. We assume a scenario in which the DRL model has already been trained by the Non-RT-RIC; therefore, operations performed during the non-real-time period (i.e., REM updates, offline model training) are not considered in this simulation.

B. Applied algorithms

In our simulations, we compare the performance of the dynamic network slicing algorithm on top of a proposed framework with static network slicing algorithm. For the static network slicing algorithm, power and bandwidth are proportionally allocated to the minimum QoS requirement in each network slice. Moreover, beam activation is randomly determined, and then users are randomly scheduled on top of signal strength map from activated beams. For the dynamic network slicing algorithm, we assume that the candidate beam is obtained from Near-RT-RIC in the proposed framework. The activation probability of each candidate beam is determined based on the number of users receiving a certain level of signal strength when the corresponding candidate beam is activated. Once beam activation is decided according to this probability, user scheduling, power allocation, and bandwidth allocation are determined by adopting Lyapunov optimization [12] to satisfy the QoS levels of users while maximizing the sum of utility functions, i.e., $U(R_k) = \log R_k$.[1]

C. Simulation Results

Fig. 2 illustrates AET and GAT for each network slice. Both results demonstrate that the dynamic network slicing algorithm on top of the proposed framework outperform static one. Moreover, the most of user throughputs in dynamic network slicing algorithm meet the QoS requirements of each network slice whereas user throughputs in static network slicing algorithm fail to meet these requirements except Ultra network slice. In addition, the dynamic network slicing algorithm shows an improvement of GAT performance with a range from 27.56% to 135.29%.

IV. Conclusion

This paper proposed a dynamic network slicing framework in conjunction with GoB beamforming. The proposed framework leverages a hierarchical structure of Open-RAN to efficiently generate dynamic network control decisions with multiple timescales in response to dynamic network environments. Then, we evaluated the proposed framework to show the superiority of the dynamic network slicing algorithm compared to the static one. As a future work, we plan to integrate deep reinforcement learning (DRL) to train deep neural network for mobility-aware beamforming and optimization to develop dynamic network slicing algorithm in a single network slicing framework.

[1]We do not provide a detailed algorithm description here since the focus of our work lies on the proposal of dynamic network slicing framework.

V. Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00225468, Development of RAN Intelligent Controller for O-RAN intelligence)

References

[1] Cisco, "Cisco annual internet report (2018-2023)," White Paper, Mar. 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf

[2] A. Filali, B. Nour, S. Cherkaoui, and A. Kobbane, "Communication and computation O-RAN resource slicing for URLLC services using deep reinforcement learning," *IEEE Communications Standards Magazine*, vol. 7, no. 1, pp. 66–73, Mar. 2023.

[3] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 836–869, Dec. 2017.

[4] M. Hoffmann and P. Kryszkiewicz, "Beam management driven by radio environment maps in O-RAN architecture," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, Rome, Italy, May 2023, pp. 54–59.

[5] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1502–1517, Dec. 2017.

[6] Y. Song and S. Xu, "Beam management based multi-cell interference suppression for millimeter wave communications," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*. Helsinki, Finland: IEEE, Jun. 2021, pp. 1–5.

[7] J. Hong, P. Yoon, S. Ahn, Y. Cho, J. Na, and J. Kwak, "Three steps toward low-complexity: Practical interference management in NOMA-based mmwave networks," *IEEE Access*, vol. 10, pp. 128 366–128 379, Dec. 2022.

[8] M. Polese, L. Bonati, S. D'oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Communications Surveys & Tutorials*, Jan. 2023.

[9] O-RAN.WG1.mMIMO-Use-Cases-TR-v01.00, "O-RAN massive MIMO use cases technical report 1.0."

[10] J. Kwak, L. Le, G. Iosifidis, K. Lee, and D. Kim, "Collaboration of network operators and cloud providers in software-controlled networks," *IEEE Network*, vol. 34, no. 5, pp. 98–105, Sep. 2020.

[11] K. Son, S. Lee, Y. Yi, and S. Chong, "REFIM: A practical interference management in heterogeneous wireless access networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 6, pp. 1260–1272, Jun. 2011.

[12] M. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, pp. 1–211, 2010.