Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HIERARCHICAL SOFTWARE-DEFINED NETWORKING ARCHITECTURE FOR NEXT-GENERATION MOBILITY MANAGEMENT

WONYONG YOON
Department of Electronics Engineering, Dong-A University, Busan, South Korea

E-mail: wyyoon@dau.ac.kr (corresponding author)

ABSTRACT

Achieving Ultra-Reliable Low-Latency Communication (URLLC) in mobility scenarios for 5G or next-generation mobile networks is challenged by the dense deployment of small cells and the high frequency of handovers. Traditional 3GPP mobility protocols and existing centralized Software-Defined Networking (SDN) architectures suffer from signaling overhead and handover delay due to relying on the top-down orchestration of forwarding path changes. To overcome this fundamental limitation, we propose an architecture of hierarchical controllers that utilize a novel bottom-up orchestration strategy for mobility support. In this bottom-up orchestration, geographically located hierarchical controllers collocated with base stations can trigger handover execution, allowing target base stations to push user plane forwarding rules up the hierarchy. Through numerical analyses, we validate that our proposed hierarchical controller architecture and bottom-up orchestration reduces network bandwidth consumption for handover signaling by 55% and handover delay by 47% compared to the traditional tunneling-based mobility management method. The results support that the proposed architecture and orchestration mechanism can be a viable solution to provide efficient mobility support in 5G or next-generation dense cellular networks.

Keywords: *Software-Defined Networking, 5G Network, Mobility, Hierarchical Controller, Bottom-Up Orchestration.*

1. INTRODUCTION

The evolution to the fifth-generation (5G) wireless networks has been architecturally defined by the need to support three distinct service classes: enhanced Mobile Broadband (eMBB), massive Machine-Type Communications (mMTC), and Ultra-Reliable Low-Latency Communication (URLLC) [1] [2]. URLLC promises to serve as the critical enabler for mission-critical applications, e.g., real-time industrial automation in smart factories and advanced vehicle-to-everything (V2X) communications for autonomous driving [3]. eMBB promises to provide high data rates for applications like virtual reality and this has driven network operators to utilize new high-frequency radios, most notably in the millimeter wave (mmWave) spectrum (e.g., 28 GHz, 39 GHz) [4]. These high-frequency signals suffer from severe propagation and penetration losses and are extremely susceptible to blockage from common obstacles. This leads to the necessity for deploying a hyper-dense network infrastructure composed of a massive number of base stations (gNBs) and small cells. However, due to the reduced range of gNBs, the number of necessary handovers increases as a User Equipment (UE) moving at even moderate pedestrian or slow vehicular speeds may triggers handover procedures far more often than in legacy 4G LTE macro-cell environments. How to support URLLC requirement in the presence of this increased mobility scenario caused by eMBB needs to be investigated.

In the standard 3GPP 5G architecture, the mobility management framework in the core network part is designed to centrally handle UE mobility, but it was not optimized for such high-frequency dense scenarios [5]. When a UE moves from a source gNB to a target gNB, the control plane of the gNBs triggers a necessary handover procedure, which will be centrally coordinated by the Access and Mobility Management Function (AMF) in the core network. A critical part of this process involves updating the user plane path, which runs via GPRS Tunneling Protocol (GTP-U) tunnels [6] from the core

network's User Plane Function (UPF) gateway to the serving gNB, to re-route traffic to the new target gNB [7]. This tunneling-based path change is a complex signaling procedure involving the UE, the source and target gNBs, and core network functions like the AMF and Session Management Function (SMF). This procedure may induce non-trivial signaling overhead and latency for each handover event, especially in the dense mmWave environment for eMBB. This signaling bottleneck may not only consume critical network bandwidth but also make it difficult to guarantee the strict latency requirements of URLLC flows during handover.

Unlike the traditional distributed routing principles in the Internet and also in IP transport networks of the 3GPP cellular networks, Software-Defined Networking (SDN) separates the control plane from the data plane of networks [8]. This decoupled architecture allows for a centralized SDN controller, with a central view of the whole network topology, to manage traffic flows and network policies in a programmable and flexible manner. Naturally SDN has emerged as one of the most promising architectural paradigm shifts to provide more flexibility in resource management in the wireless cellular networks [9]. Specifically, a centralized controller with connections to base stations deals with all necessary decisions involved in handover like handover signaling, target cell selection, and admission control in order to handle frequent handovers inherent in ultra-dense 5G networks [10]. Shah et al. [11] proposed the integration of SDN and multiaccess edge computing (MEC) for latency-sensitive vehicular users in 5G networks, providing seamless coverage and service continuity to the mobile users. An SDN controller centrally connects with each gNB and each UPF for top-down orchestration of mobility processes. Zeng et al. [12] propose that a central SDN controller allows for mobility-aware proactive flow setup in the mobile edge for latency-sensitive applications to minimize handover delay by establishing user plane paths before the handover actually occurs.

While the initial integration of SDN principles into the 5G networks provided the necessary architectural flexibility and programmability, relying solely on a centralized orchestration model introduced inherent latency and scalability issues particularly for frequent mobility events [13]. Incorporating multiple SDN controllers in a distributed fashion has been targeted for this issue [14]. Kim et al. [15] propose a two-level hierarchy of SDN controllers for enhancement in which controllers are involved in separate edge clouds and the central cloud. However, a handover procedure requires the controller in the central cloud to select a target edge cloud, and thereby resulting in top-down orchestration of forwarding rule changes. ES-5G model [16] embeds SDN control closer to the edge access network in which distributed controllers are in charge of their corresponding edge and share topology information to construct an entire network topology. These efforts moved control functions from a purely centralized entity to a locally distributed network architecture but still rely on traditional top-down orchestration of forwarding rules.

In this paper, we notice that some points of mobility procedures proposed by the prior studies are open for further enhancement in terms of network bandwidth consumption and handover delay. We take an SDN-based approach to address the issue and propose a novel architecture and new procedures to support mobility with lesser network overhead and lower latency in 5G networks. In doing so we also make a fundamental departure from the centralized top-down orchestration of a single centralized SDN controller or multiple flat controllers, and rather we propose a new structure of hierarchically organized SDN controllers and *bottom-up orchestration* of forwarding path modification during handover.

The remainder of the paper is organized as follows. Section 2 discusses two generic approaches to support mobility management in 5G cellular networks. Section 3 proposes a novel hierarchical controller-based architecture and relevant handover procedures. Section 4 presents numerical analysis and results for comparison purposes. Section 5 summarizes previous related works. Section 6 provides our concluding remarks.

2. TWO APPROACHES FOR 5G MOBILITY MANAGEMENT

In this section we discuss two approaches to support user equipment (UE) mobility in 5G wireless networks. One is the traditional 3GPP standard-based mobility management and the other is SDN-based mobility management. In Figure 1, the 5G access and core network architecture is depicted as specified in the 3GPP standard [5]. The 3GPP 5G architecture fundamentally overhauls mobility management by separating the control plane and user plane within its new Service-Based Architecture (SBA) core [5] [7]. Mobility management is primarily centralized in the Access and Mobility

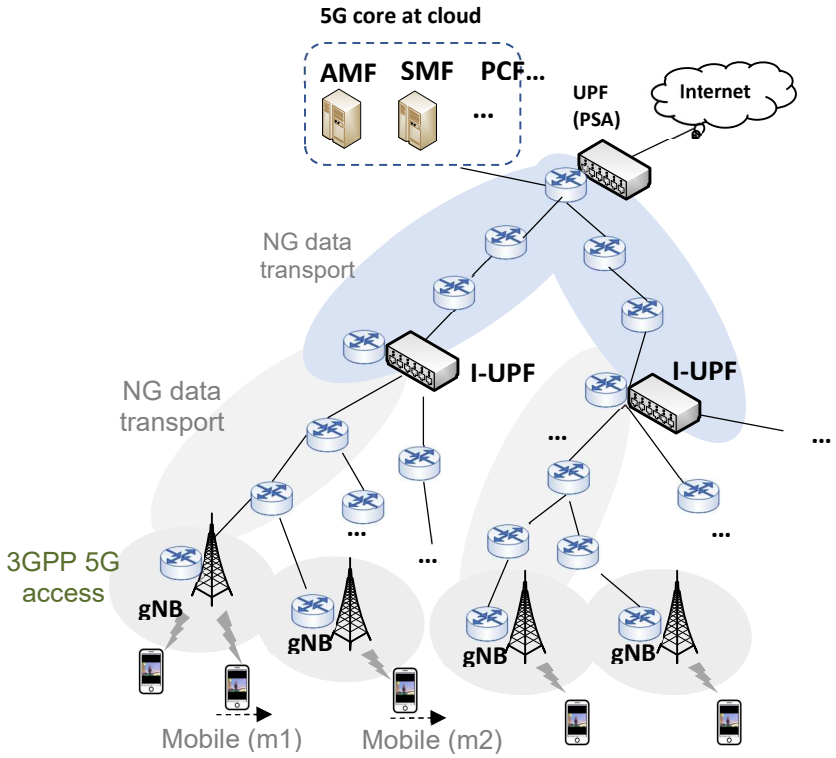

Figure 1: The 5G access and core network architecture.

Management Function (AMF), which is responsible for handling UE registration, connection status, and reachability. UEs in an actively connected state necessitate seamless handovers to maintain service continuity for services like URLLC. In the 5G core, the Session Management Function (SMF) establishes the UE's PDU session, which defines the user plane path, while the User Plane Function (UPF) acts as the data anchor and gateway. This user plane path is physically realized as a GTP-U tunnel between the UPF and the serving gNB (base station). During a handover, the AMF coordinates the context transfer, and the SMF instructs the UPF to update its tunneling endpoint, switching the data path from the source gNB to the target gNB. In Xn-based handover, two gNBs can manage the handover directly over the Xn interface with minimal core network involvement, which is crucial for reducing interruption time. We note that reliance on a tunneling-based framework for path switching remains a source of overhead and latency, particularly in dense networks with frequent mobility events.

The traditional non-SDN approach uses tunneling mechanisms to support mobility [17], that is, tunnels are created and removed depending on the location of UE. A handover procedure is performed when a mobile changes its contacting gNB without the change of the upward I-UPF (mobile m1 in Figure 1) or a mobile changes its gNB along with the change of the upward I-UPF (mobile m2 in Figure). In either case, the creation and removal of tunnels are centrally coordinated by AMF and SMF in the 5G core side. Routing between two endpoints of a tunnel is done by distributed routing protocols, e.g., OSPF, among in-between routers/switches. Figure 2 describes the handover procedure for the former mobility scenario while Figure 3 depicts the one for the latter mobility scenario.

We first explain each step in Figure 2 for the mobility scenario without I-UPF replacement. In step 1, the UE reports its signal measurement results according to the configuration by its current gNB [18]. In step 2, the gNB makes a decision on handover of the UE depending on measurement results and then sends Handover Request message to the new target gNB. In step 3, the target gNB performs admission control and then responses with Handover Request Acknowledgment which contains a RRCReconfiguration message. In step 4, the source

Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

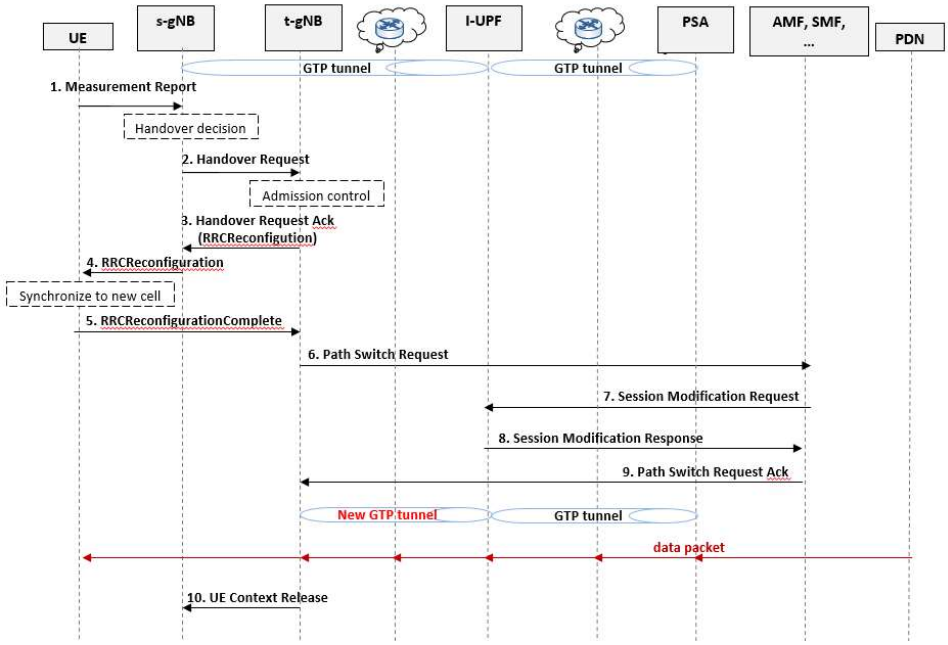

Figure 2: The handover procedure without I-UPF change in the 3GPP tunneling-based mobility management.

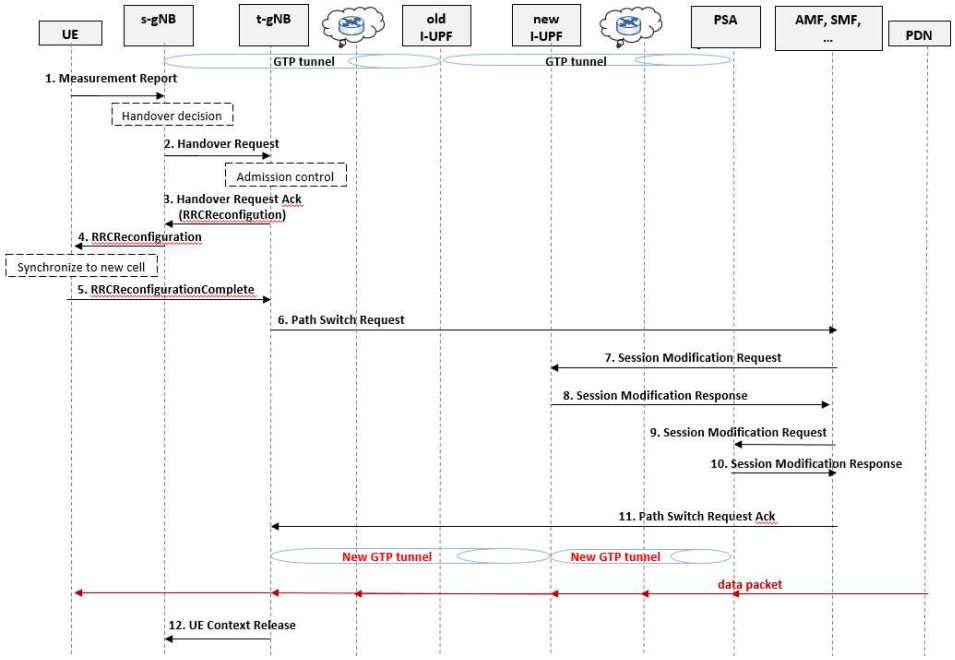

Figure 3: The handover procedure with the change of I-UPF in the 3GPP tunneling-based mobility management.

gNB sends the RRCReconfiguration message to the UE to trigger camping on a new target cell. In step 5, once synchronized with the new cell, the UE sends RRCReconfigurationComplete to the target gNB. In step 6, the target gNB needs to open a new tunnel with the upward I-UPF and therefore sends Path Switch Request containing AN tunnel information of the target gNB to the AMF in the core cloud via N2 or NGAP [19]. The AMF/SMF in the cloud now centrally coordinates the I-UPF and the target gNB so that they can make a necessary tunnel. Since we assume the AMF and the SMF are located in the core cloud, messages between them seem missing in the procedure. In step 7, the AMF sends Session Modification Request via N4 interface. In step 8, the I-UPF sends Session Modification Response

containing CN tunnel information of the I-UPF to the AMF. In step 9, the AMF sends Path Switch Request Acknowledgment and then the target gNB gets aware of the CN tunnel information of the I-UPF. At this point a new tunnel between the target gNB and the I-UPF is established and data packets can be transmitted over an end-to-end path. In step 10, the target gNB sends UE Context Release to the source gNB.

Now we explain each step in Figure 3 for the mobility scenario with I-UPF replacement. The radio part of the procedure in step 1 through 5 is same as above. In step 6, the target gNB needs to open a new tunnel with a new I-UPF and therefore sends Path Switch Request containing AN tunnel information of the target gNB to the AMF in the core cloud via N2 or NGAP. Since the I-UPF is new to the end-to-end path, the AMF now needs to centrally coordinate a new tunnel between the I-UPF and the target gNB and also a new tunnel between the I-UPF and the target gNB. Since we assume the AMF and the SMF are located in the core cloud, messages between them seem missing in the procedure. In step 7, the AMF sends Session Modification Request to the new I-UPF via N4 interface. In step 8, the new I-UPF sends Session Modification Response containing CN tunnel information of the I-UPF to the AMF. In step 9, the AMF sends Session Modification Request to the PSA. In step 10, the PSA sends Session Modification Response containing CN tunnel information of the PSA to the AMF. In step 11, the AMF sends Path Switch Request Acknowledgment and then the target gNB gets aware of the CN tunnel information of the I-UPF. At this point a new tunnel between the target gNB and the I-UPF and a new tunnel between the PSA and the I-UPF are established and data packets can be transmitted over an end-to-end path. In step 12, the target gNB sends UE Context Release to the source gNB.

Meanwhile, in the SDN-based approach, typically a single SDN controller resides at the core cloud and coordinates the forwarding functions of the entire switches/routers between the PSA and gNBs depending on the location of UEs. All forwarding entities have one-to-one TCP connection with a single SDN controller to communicate directly and thus they are considered belonging to a flat architecture. During handover, a target gNB send a path modification for its incoming mobile to the SDN controller, which with the global view of the entire network topology calculates a new routing path from the PSA to the target gNB and orders necessary forwarding rules to the switches on the new path. In this approach I-UPF is considered as a flat switch by the controller and thus the handover procedures with/without the change of I-UPF remain same.

We note that the both approaches operate in a top-down manner. In the traditional non-SDN approach, the AMF at the core side coordinates the tunneling mechanism of the PSA, I-UPFs, and gNBs in a top-down manner. In the SDN-based approaches, the central SDN controller orchestrates the forwarding mechanism ad all forwarding entities including the PSA, I-UPFs, gNBs, and all in-between transport switches. The handover triggering entity is the target gNB which sends a request message to the central core entity (either the AMF or the SDN controller). The central entity sends control messages to relevant forwarding entities. This kind of central control in a top-down way causes latency and network bandwidth consumption during handover. This motivates us to propose a bottom-up coordination mechanism to reduce latency and network bandwidth consumption in 5G mobility management.

3. THE PROPOSED ARCHITECTURE

In this section we propose a novel hierarchical architecture in which SDN controllers collocated at gNBs, I-UPFs, and the PSA provide the bottom-up orchestration of forwarding rules to support mobility in a faster and more bandwidth-efficient way. Distributed mobility applications running on the hierarchical SDN controllers communicate with each other through the east-west interface to complete the handover procedure in a collaborative way. Figure 4 depicts the proposed architecture. The top controller collocated with the PSA (the gateway to/from the Internet) orchestrates downward I-UPFs and in-between switches/routers. The controllers collocated with I-UPFs controls gNBs and in-between switches/routers. If the I-UPF have downward I-UPFs then it controls those downward I-UPFs and in-between switches. The controller at a gNB orchestrates its upward I-UPF and switches along the upward routing path. By decomposing the control area of the entire network to segmentations orchestrated by hierarchical controllers, the proposed architecture can limit handover-induced message exchanges only within necessary areas and thus reduce the handover latency and handover network bandwidth overhead. The coordination of forwarding rules can be triggered in a bottom-up manner from gNBs where actual UE mobility

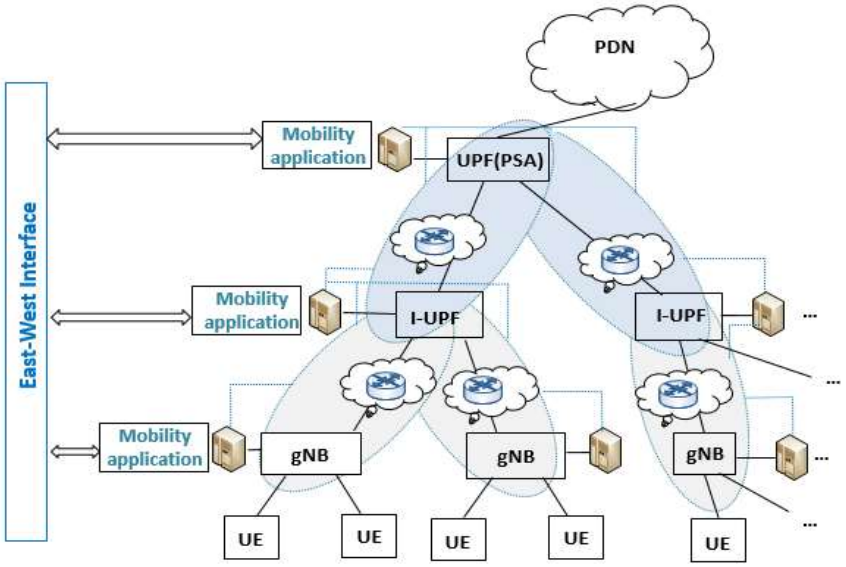

Figure 4: The propose hierarchical architecture.

happens. This is fundamental departure from the previous works in which the core network entity far from gNBs triggers mobility procedures in a top-down manner.

Figure 5 illustrates the handover procedure of the proposed architecture for a mobility scenario without I-UPF relocation. The radio network part is same as the 3GPP standard (step 1 through 5). At step 6, the SDN controller at the target gNB calculates a new path between the target gNB and the current I-UPF and sets up required forwarding rules at in-between switches/routers. This is called bottom-up orchestration. All required rules are included in one Rule Change message, which is transmitted along the new path toward the I-UPF SDN controller. Any in-between switch that finds out engaged forwarding rules in the Rule Change message configures its forwarding table accordingly and then resends the Rule Change message upward. The I-UPF SDN controller that receives the Rule Change message needs to only synchronize the new path and relevant forwarding rules. Now the new path is activated and thus packets from the PSA to the UE begins to traverse the new path between the I-UPF and the target gNB. Note that the handover delay is reduced due to bottom-up orchestration. This reconfiguration impacts the I-UPF and downward switches only without further change in upward switches. Finally, the I-UPF controller sends the notification of mobile location change to the top SDN controller, which then records the up-to-date location of the mobile and sends an acknowledgment to the target gNB SDN controller.

Figure 6 illustrates the handover procedure of the proposed architecture for a mobility scenario with I-UPF relocation. The radio network part is same as the 3GPP standard (step 1 through 5). At step 6, the SDN controller at the target gNB calculates a new path between the target gNB and the new I-UPF and sets up required forwarding rules at in-between switches/routers. All required rules are included in one Rule Change message, which is transmitted along the new path toward the new I-UPF SDN controller. Any in-between switch that finds out engaged forwarding rules in the Rule Change message configures its forwarding table accordingly and then resends the Rule Change message upward. The I-UPF SDN controller that receives the Rule Change message needs to synchronize the new path and relevant forwarding rules. The SDN controller at the new I-UPF calculates a new path between itself and the PSA and sets up required forwarding rules at in-between switches/routers. All required rules are included in one Rule Change message, which is transmitted along the new path toward the PSA-collocated controller. Any in-between switch that finds out engaged forwarding rules in the Rule Change message configures its forwarding table accordingly and then resends the Rule Change message upward. Now the new path is activated and thus packets from the PSA to the UE begins to traverse the new path between the PSA and the target gNB. Note that the handover delay is reduced due to bottom-up orchestration. Finally, the PSA controller sends the notification of mobile location change to the top SDN controller, which then records the up-

Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

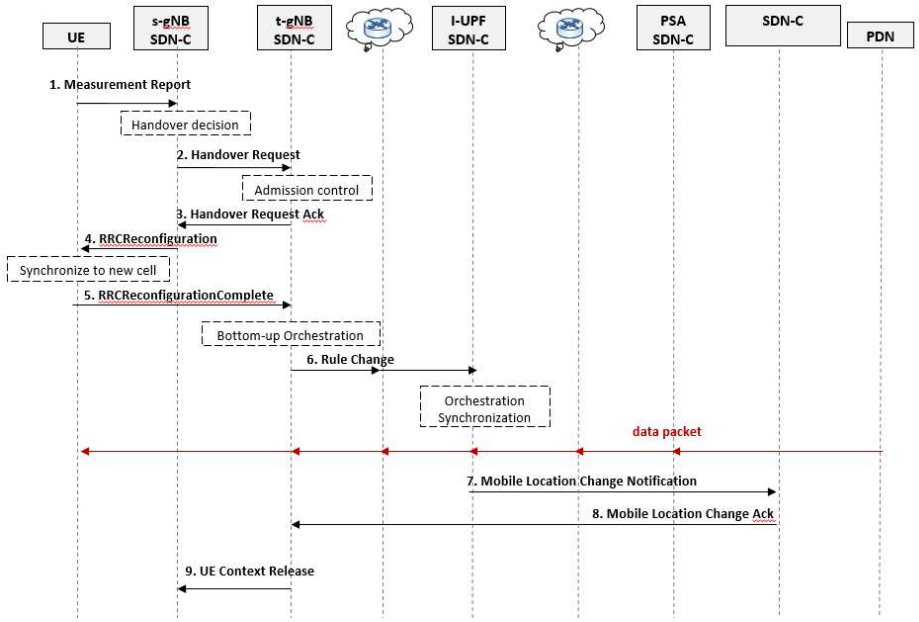

Figure 5: The handover procedure without the change of I-UPF in the hierarchical SDN-based approach.

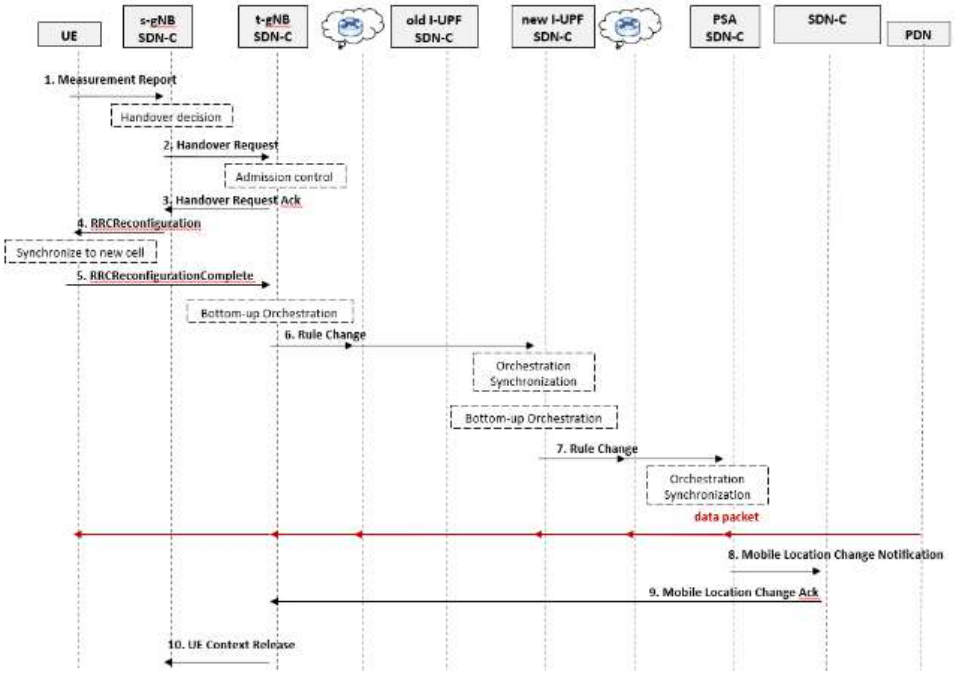

Figure 6: The handover procedure with the change of I-UPF in the hierarchical SDN-based approach.

to-date location of the mobile and sends an acknowledgment to the target gNB SDN controller.

4. PERFORMANCE ANALYSIS

In this section, we present performance analyses and results for the purpose of comparison of 3GPP tunneling based mobility management and the proposed hierarchical controller based mobility management. We consider both the control plane and the data plane for two Xn-based handover scenarios with and without UPF replacement.

4.1 Control plane analysis

We consider two different handover scenarios, i.e., Xn-based handover without UPF replacement (scenario 1) and Xn-based handover with UPF

replacement (scenario 2). We first look into network bandwidth consumption incurred by control messages of mobility procedures. We focus on control messages between the core network and the access network but do not consider those between gNBs because they are local message exchange. We let B_1^T and B_2^T denote network bandwidth consumption of 3GPP tunneling-based architecture for scenario 1 and scenario 2, respectively. We let B_1^H and B_2^H denote network bandwidth consumption of the proposed hierarchical controller based architecture for scenario 1 and scenario 2, respectively.

For the purpose of simplicity but without loss of generality, we assume that the target cellular network topology is k-ary tree rooted at PSA (PDU Session Anchor) which is the top UPF. I-UPF nodes are located at distance p (path length) from their parent UPF. The hierarchical depth of UPF nodes is called height and the maximum height is denoted by H. We assume that gNBs are located at distance p from their belonging UPF. Therefore, the total number of gNBs at height h is calculated as M^h where $M = k^p$. With parameter n (the number of mobile terminals per gNB) and r (handover rate), the mean of the number of handover procedures X triggered per gNB can be derived in the model of binomial probability distribution as follows.

$$\mathrm{E}[X] = \sum_{i=0}^{n} i \binom{n}{i} r^i (1-r)^{n-i} = nr$$

Therefore, the total number of handover procedures triggered at gNB height h is $nrM^h = nrk^{ph}$. For scenario 1, each handover procedure includes the request-response pair between AMF and gNB along total $h \cdot p$ links plus p links and the request-response pair between AMF and current I-UPF along total (h-1)$\cdot p$ links plus p links. The whole network bandwidth consumption by all gNBs is derived in the following equation.

$$B_1^T = \sum_{h=1}^{H} nrk^{ph}(4hp + 2p)$$

For scenario 2, each handover procedure includes the request-response pair between AMF and gNB along total $h \cdot p$ links plus p links and the request-response pair between AMF and new I-UPF along total (h-1)$\cdot p$ links plus p links. We assume that the request-response pair between AMF in the cloud and PSA takes total p links to set up the tunnel between PSA and new I-UPF. The whole network bandwidth consumption by all gNBs is derived in the following equation.

$$B_2^T = \sum_{h=1}^{H} nrk^{ph}(4hp + 4p)$$

Now we analyze network bandwidth consumption by the proposed hierarchical controller architecture. For scenario 1, each handover procedure includes the rule installment message sent by the SDN controller at the target gNB to the controller at the current I-UPF along p links in one way direction, and a message by the controller at the I-UPF to AMF in the cloud to notify the new cell information of the mobile terminal along total (h-1)$\cdot p$ links in one way direction. The whole network bandwidth consumption by all gNBs is derived in the following equation.

$$B_1^H = \sum_{h=1}^{H} nrk^{ph}(2hp)$$

For scenario 2, each handover procedure includes the rule installment message sent by the controller at the target gNB to the controller at the new I-UPF along p links in one way direction, the rule installment message sent by the controller at the new I-UPF to the controller at PSA along (h-1)$\cdot p$ links in one way direction, and a message by the controller at PSA to AMF in the cloud to notify the new cell information of the mobile terminal along total p links in one way direction. The whole network bandwidth consumption by all gNBs is derived in the following equation.

$$B_2^H = \sum_{h=1}^{H} nrk^{ph}(2hp + 2p)$$

Now we turn our focus to handover delay during mobility procedures. We let D_1^T and D_2^T denote the handover delay of 3GPP tunneling-based architecture for scenario 1 and scenario 2, respectively. We let D_1^H and D_2^H denote the handover delay of the proposed hierarchical controller-based architecture for scenario 1 and scenario 2, respectively.

For scenario 1, each handover procedure for 3GPP tunneling-based architecture takes delay in the message sequence chart in Figure. The delay for setting up Xn tunnel between a previous gNB and a new gNB takes two message exchanges over a wired

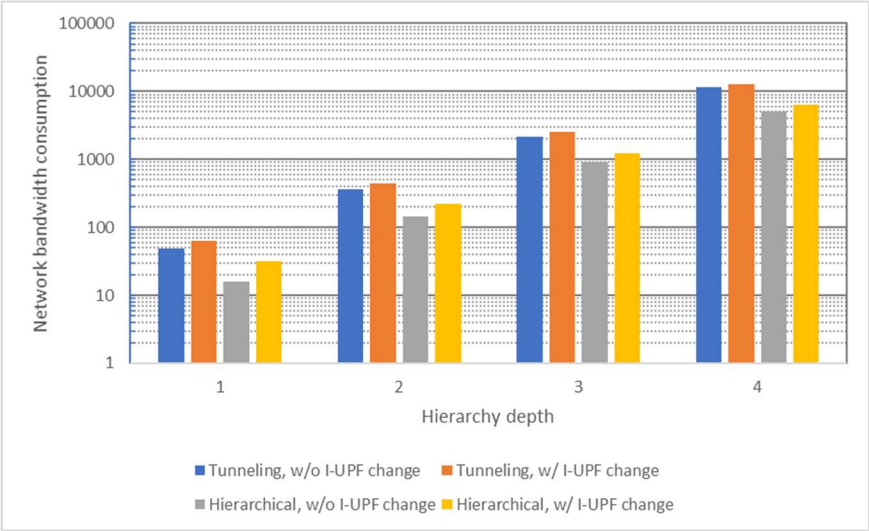

Figure 7: The comparison of network bandwidth consumption for mobility scenarios.

link and the delay for sending handover command to a mobile terminal and camping on the new gNB takes two message exchanges over the radio channel. This adds $2d_{wired} + 2d_{radio}$ to the total handover delay. The procedure for setting up new tunnels in upward and downward direction between current I-UPF and new gNB includes the request-response pair between gNB and AMF and the request-response pair between AMF and I-UPF. The request-response pair between AMF and gNB needs transmission along total $h \cdot p$ links plus p links and thus the delay incurred is calculated as the multiplication of the number of traversed links and transmission delay per wired link, which is $2(hp + p) \cdot d_{wired}$. The request-response pair between AMF and current I-UPF along total (h-1)·p links plus p links incurs delay which amounts to $2((h-1)p + p) \cdot d_{wired}$. Thus, the total handover delay for each handover without I-UPF replacement is obtained in the following equation.

$$D_1^T = (4hp + 2p + 1) \cdot d_{wired} + 2d_{radio}$$

For scenario 2, each handover procedure for 3GPP tunneling-based architecture takes delay in the message sequence chart in Figure. The delay for setting up Xn tunnel between a previous gNB and a new gNB takes two message exchanges over a wired link and the delay for sending handover command to a mobile terminal and camping on the new gNB takes two message exchanges over the radio channel. This adds $2d_{wired} + 2d_{radio}$ to the total handover delay. The procedure for setting up new tunnels in upward and downward direction between new I-UPF and new gNB and also that for setting up new tunnels in upward and downward direction between new I-UPF and PSA are considered in handover delay computation. The procedure for setting the both tunnels includes the request-response pair between new gNB and AMF, the request-response pair between AMF and new I-UPF, and the request-response pair between AMF and PSA. The request-response pair between gNB and AMF needs transmission along total $h \cdot p$ links plus p links and thus the delay incurred is calculated as the multiplication of the number of traversed links and transmission delay per wired link, which is $2(hp + p) \cdot d_{wired}$. The request-response pair between AMF and new I-UPF along total (h-1)·p links plus p links incurs delay which amounts to $2((h-1)p + p) \cdot d_{wired}$. The request-response pair between AMF and PSA along p links incurs delay $2p \cdot d_{wired}$. Thus, the total handover delay for each handover with I-UPF replacement is obtained in the following equation.

$$D_2^T = (4hp + 4p + 1) \cdot d_{wired} + 2d_{radio}$$

For scenario 1, each handover procedure for the proposed hierarchical controller-based architecture takes delay in the message sequence chart in Figure. The delay for setting up forwarding rules between a previous gNB and a new gNB takes two message exchanges over a wired link and the delay for sending handover command to a mobile terminal and camping on the new gNB takes two message exchanges over the radio channel. This adds $2d_{wired} + 2d_{radio}$ to the total handover delay. each handover procedure includes the rule installment message sent by the controller at the target gNB to the controller at the current I-UPF along p links in

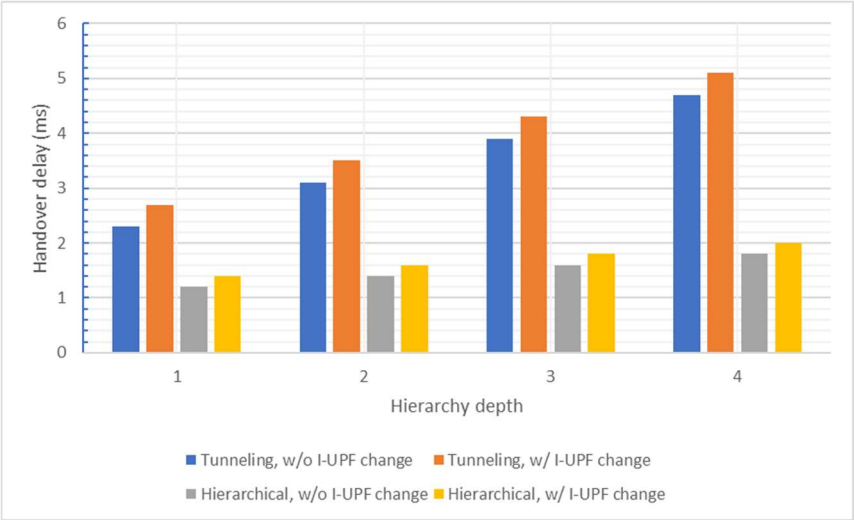

Figure 8: The comparison of handover delay for mobility scenarios.

one way direction, and a message by the controller at the I-UPF to AMF in the cloud to notify the new cell information of the mobile terminal along total $(h\text{-}1)\cdot p$ links in one way direction. Thus, the total handover delay for each handover without I-UPF replacement is obtained in the following equation.

$$D_1^H = (hp) \cdot d_{wired} + 2d_{radio}$$

For scenario 2, each handover procedure for the proposed hierarchical controller-based architecture takes delay in the message sequence chart in Figure. The delay for setting up forwarding rules between a previous gNB and a new gNB is obtained in the same manner as that for scenario 1. Each handover procedure includes the rule installment message sent by the controller at the target gNB to the controller at the current I-UPF along p links in one way direction, a message by the controller at the I-UPF to AMF in the cloud to notify the new cell information of the mobile terminal along total $(h\text{-}1)\cdot p$ links in one way direction, and a message by the controller at PSA to AMF in the cloud to notify the new cell information of the mobile terminal along total p links in one way direction. Thus, the total handover delay for each handover with I-UPF replacement is obtained in the following equation.

$$D_2^H = (hp + p) \cdot d_{wired} + 2d_{radio}$$

We plot the network bandwidth consumption for different mobility scenarios in Figure 7. The horizontal axis represents the depth of UPF hierarch in a given wireless network. The vertical axis represents the network bandwidth consumption in the entire wireless network. Note that the number of handover increases more than exponentially as the hierarch depth increases the network size exponentially. We set the number of mobile nodes to 100 and the handover rate to 0.01 for simplicity. We may also set different numbers for them. The network bandwidth consumption is measured by the number of signaling messages passing links in the whole network. The message reduction ratio due to the proposed bottom-up orchestration rises up to about 55 percent. We also plot the handover delay for different mobility scenarios in Figure 8. The vertical axis represents the handover delay per handover event. The proposed bottom-up orchestration mechanism contributes to reducing the delay up to almost 47 percent.

4.2 Data plane analysis

In this section we compare the header overhead of data plane packets for 3GPP tunneling-based architecture [20] [21] and the proposed hierarchical controller-based architecture. In 3GPP tunneling-based architecture, data plane packets are transmitted through tunnels between gNB and I-UPF (N3 reference point), tunnels between I-UPF and I-UPF (N9 reference point), and tunnels between I-UPF and PSA (N9 reference point). A normal IP packet is encapsulated by a sending tunnel endpoint node into another IP packet with IPv6/UDP/GTP-U header outside, which is decapsulated by a receiving tunnel endpoint node. GTP-U defined in 3GPP standard TS 29.281 [6] requires 16-byte GTP-U header for each data plane packet which includes 4-byte TE-ID (Tunnel Endpoint ID) of the receiver side of the packet. UDP header length is 8 bytes which include source port number, destination port number, length, and checksum field. Outer IPv6 header length is fixed to 40 bytes which include version, priority, flow label, payload length, next

header, hop limit fields along with 128-bit source address and 128-bit destination address. The total header overhead for tunneling is 64 bytes per data packet. This results in reduction in goodput in the data plane by 4.26%.

4.3 Discussion

The results of the conducted numerical analyses imply that the proposed hierarchical architecture and bottom-up mobility orchestration improves key performance measures like network overhead and handover delay. The results may open new research problems and issues for future cellular and mobile networks in various aspects of not just connected-state handover but also initial attach, idle-state mobility, roaming procedures, etc. as our work is a departure from traditional centralized architecture.

5. RELATED WORK

The most concerned work to our paper is ES-5G proposed by Chiang and Shang [16], which is a novel 5G architecture with multiple distributed SDN controllers to minimize end-to-end latency. The SDN controllers orchestrate the data plane of UPFs and gNBs while the control plane still interfaces with AMF in the central cloud core network. The SDN controllers directly place forwarding rules on UPFs, gNBs, and switches/routers, which eliminates the need for tunneling-based forwarding. The controllers share their topology information to make and share global view of the entire network topology. The SDN controllers are organized in a flat structure while our proposed architecture is hierarchical in nature and information sharing is limited within regions. Their SDN controllers perform the top-down orchestration of forwarding path changes during handover which has room for enhancement in procedures. In contrast to it, our proposed one enables the bottom-up orchestration during handover to reduce handover delay and network bandwidth consumption.

Abdulghaffar et al. [13] propose a new architecture in which a single centralized SDN controller interacts with the 5G core by the northbound interface, and with individual forwarding entities like UPF and gNB by the southbound API. The authors develop analytical models to represent the behavior and performance of their proposed SDN-based 5GC. demonstrating They validate that the new SDN-based core is a viable solution for enabling significant reductions in signaling overhead and improved resource utilization. The difference from our work is the use of the single centralized controller and the overhead of tunneling-based forwarding. The installment of packet forwarding rules on IP transport routers/switches is not considered in the SDN realm so existing distributed routing protocols are still needed even in the presence of a SDN controller.

Kim et al. [15] introduce a two-level hierarchy of SDN enhancement to the existing 4G LTE network architecture. The mobile core functionalities are virtualized into the edge and central cloud. The forwarding plane functions (such as S-GW, P-GW, and MME) in charge of bearer delivery and data service are located at the edge cloud. The others, such as user and service management in charge of policy and control functions (such as HSS, PCRF, and MME), are included in the central cloud. The mobility management of UE is mainly performed by MME in the edge cloud, while the MME in the central cloud has a location management role in the UE handover between edge clouds. The UE movement in the same edge cloud area can be easily handled by the local SDN controller. The decision for data forwarding is taken in the control plane in each selected edge cloud for supporting the required services. They still need the top-down orchestration for mobility management in which the SDN controller in the central could selects a target edge cloud. Meanwhile our proposed architecture allows for the bottom-up orchestration for mobility management to enhance handover performance.

Tadros et al. [14] propose Logically Centralized-Physically Distributed (LC-PD) architecture with multiple controllers that operate different services like Web Surfing, Video streaming, VoIP, separately. The controllers are not aimed at providing mobility management and optimal utilization of network resources. The architecture is designed not for a 3GPP inherent cellular network but for general networks and therefore it is not clear how to apply the proposed architecture to 5G networks.

Moradi et al. [22] introduce SoftBox, that redesigns the LTE cellular network core to build efficient and low latency services on a per-UE basis. The core network, the transport fabric, the RAN are all orchestrated by global SDN/NFV controller. Agents include per-UE containers for optimization and communicate with the global controller. Softbox can reduce the overhead and latency caused by GTP tunneling mechanisms of LTE. However, the design is based on the single global SDN and NFV controller which induces inherent latency and inefficiency of network bandwidth.

Alotaibi and Nayak [23] investigate the fundamental tradeoff between handover delay and SDN controller load balancing in heterogeneous networks with a hierarchy of SDN controllers. The algorithm's core idea is to link the handover process directly to the real-time load status of all available access points, not just signal strength. The primary contribution is this integrated approach, which simultaneously optimizes for low latency (by ensuring a smooth transition) and efficient load distribution (by avoiding congested access points). However, the work is not tailored for 5G cellular networks but for general networks. It is not clear how the work will be applied in 5G cellular networks in an efficient way.

In summary, these prior works have been proposed for top-down orchestration of mobility support either in a centralized controller architecture or in a distributed controller architecture. Our proposed work differs from these prior works in that its novel bottom-up orchestration of mobility support can result in lesser network overhead and lower latency in 5G and beyond networks.

6. CONCLUSION

In this paper, we presented a novel bottom-up orchestration mechanism in the hierarchical controller architecture for 5G wireless networks. The bottom-up orchestration, compared to traditional top-down control, enables lesser latency and less network overhead for 5G mobility management since the handover procedure is initiated nearby mobility happens. Through numerical analyses in the varied range of scalability, we demonstrated that the novel bottom-up orchestration can contribute to reduce network bandwidth consumption for handover signaling by 55% and handover delay by 47% compared to the traditional tunneling-based mobility management method. Our work makes a significant addition to literature in that a new bottom-up approach to mobility management is firstly proposed in contrast to traditional top-down approaches.

Depending on network deployment, there can be mobility scenarios where Xn interface between gNBs is not available and thus N2-based handover is needed. We plan to extend the current work to N2-based handover scenarios and confirm that the proposed novel bottom orchestration will work in a more efficient manner. We plan to investigate the applicability of proposed bottom-up SDN orchestration in the next-generation 6G cellular networks as well. Another direction for enhancing the current work is to implement the proposed mobility management mechanism in open-source SDN controller codes and integrate the distributed controllers with a 5G network simulator. This work will demonstrate the feasibility and performance of the proposed mechanism.

ACKNOWLEDGMENT:

This work was supported by the Dong-A University Research Fund. The corresponding author is W. Yoon.

REFRENCES:

[1] Nhu-Ngoc Dao, Ngo Hoang Tu, Trong-Dai Hoang, Tri-Hai Nguyen, Luong Vuong Nguyen, Kyungchun Lee, Laihyuk Park, Woongsoo Na, and Sungrae Cho, “A review on new technologies in 3GPP standards for 5G access and beyond,” *Computer Networks*, Vol. 245, May 2024.

[2] M. E. Haque, F. Tariq, M. R. A. Khandaker, K. -K. Wong, and Y. Zhang, "A Survey of Scheduling in 5G URLLC and Outlook for Emerging 6G Systems," *IEEE Access*, Vol. 11, April 2023 pp. 34372-34396.

[3] B. S. Khan, S. Jangsher, A. Ahmed, and A. Al-Dweik, "URLLC and eMBB in 5G Industrial IoT: A Survey," *IEEE Open Journal of the Communications Society*, Vol. 3, July 2022, pp. 1134-1163.

[4] N. Saba, P. Lassila, K. Ruttik, R. Jäntti, and J. Salo, "Radio Network Planning for 5G FWA at 3.5 GHz and 26 GHz: A Link- and Flow-Level Approach," *IEEE Access*, Vol. 13, August 2025, pp. 152782-152799.

[5] TS23.501 "System Architecture for the 5G System; Stage 2". 3GPP Release 16.20.0, June 2024.

[6] TS 29.281, “General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U),” 3GPP Release 17.2.0, May 2022.

[7] TS23.502 “Procedures for the 5G System (5GS); Stage 2,” 3GPP Release 16.19.0, March 2024.

[8] B. Raghavan, M. Casado, T. Koponen, S. Ratnasamy, A. Ghodsi, and S. Shenker, "Software-defined Internet architecture: Decoupling architecture from infrastructure," *Proc. 11th ACM Workshop Hot Topics Netw. (HotNets-X)*, October 2012, pp. 43-48.

[9] Z. Zaidi, V. Friderikos, Z. Yousaf, S. Fletcher, M. Dohler, and H. Aghvami, "Will SDN Be Part of

5G?," *IEEE Communications Surveys & Tutorials*, Vol. 20, No. 4, Fourth quarter 2018, pp. 3220-3258.

[10] T. Bilen, B. Canberk and K. R. Chowdhury, "Handover Management in Software-Defined Ultra-Dense 5G Networks," *IEEE Network*, Vol. 31, No. 4, July-August 2017, pp. 49-55.

[11] S. D. A. Shah, M. A. Gregory, S. Li, R. d. R. Fontes, and L. Hou, "SDN-Based Service Mobility Management in MEC-Enabled 5G and Beyond Vehicular Networks," *IEEE Internet of Things*, Vol. 9, No. 15, August 2022, pp. 13425-13442.

[12] Y. Zeng et al., "Mobility-Aware Proactive Flow Setup in Software-Defined Mobile Edge Networks," *IEEE Transactions on Communications*, Vol. 71, No. 3, March 2023, pp. 1549-1563.

[13] Abdulaziz Abdulghaffar, et al. "Modeling and Evaluation of Software Defined Networking Based 5G Core Network Architecture," *IEEE Access*, January 2021, pp. 10179-10198.

[14] Catherine Nayer Tadros, Mohamed R. M. Rizk, and Bassem Mahmoud Mokhtar, "Software Defined Network-Based Management for Enhanced 5G Network Services," *IEEE Access*, March 2020, pp. 53997-54008.

[15] Yong-hwan Kim, Joon-Min Gil, and Dongkyun Kim, "A location-aware network virtualization and reconfiguration for 5G core network based on SDN and NFV," *International Journal of Communication System*, January 2021, pp. 1099-1131.

[16] Wei-Kuo Chiang and Yi-Hsin Shang, "ES-5G: A Novel Edge-based SDN-enabled 5G Architecture for Lower Latency," *6th International Conference on Information Science and Systems, International Conference on Information Science and Systems(ICISS)*, November 2023, pp. 143-153.

[17] M. Tayyab, X. Gelabert, and R. Jäntti, "A Survey on Handover Management: From LTE to NR", *IEEE Access*, Vol. 7, August 2019, pp. 118907-118930.

[18] TS38.300, "NR and NG-RAN Overall description; Stage-2," 3GPP Release 17.0.0, May 2022

[19] TS38.413, "NG Application Protocol (NGAP)," 3GPP Release 17.0.0, May 2022

[20] Wen-Long Chin, Yen-Chun Huang, Pin-An Pan, Yu-Xiang Huang, Cheng-Hsien Yu, and Hsiao-Hwa Chen, "Traffic Management for Network Slicing in P4 Data Plane with GPRS Tunneling Protocol and Two-Stage Pipeline Architecture," *IEEE Communications Magazine*, July 2025, pp. 1-8.

[21] D. Pineda, R. Harrilal-Parchment, K. Akkaya, and A. Perez-Pons, "SDN-based GTP-U Traffic Analysis for 5G Networks," *IEEE/IFIP Network Operations and Management Symposium*, June 2023, pp. 1-4.

[22] M. Moradi, Y. Lin, Z. M. Mao, S. Sen and O. Spatscheck, "SoftBox: A Customizable, Low-Latency, and Scalable 5G Core Network Architecture," *IEEE Journal on Selected Areas in Communications*, Vol. 36, No. 3, March 2018, pp. 438-456.

[23] M. Alotaibi and A. Nayak, "Linking handover delay to load balancing in SDN-based heterogeneous networks," *Computer Communications*, vol. 173, May 2021, pp. 170-182.