VirtRAN : An SDN/NFV based Framework for 5G RAN Slicing

Akshatha Nayak Manjeshwar[1], Pranav Jha[1], Abhay Karandikar[1,2], and Prasanna Chaporkar[1]

[1]Department of Electrical Engineering, Indian Institute Technology Bombay
Email: {akshatha, pranavjha, karandi, chaporkar}@ee.iitb.ac.in
[2]Director, Indian Institute Technology Kanpur
Email: {karandi}@iitk.ac.in

Abstract

The upcoming Fifth Generation (5G) mobile network aims to support a wide variety of services. In addition to the four standardized service use cases, network operators are also looking for the ability to deploy newer services in shorter timescales in order to quickly monetize the 5G network. This has resulted in the emergence of Software Defined Networking (SDN) and Network Function Virtualization (NFV) as key technologies for designing the 5G networks. In this paper, we provide a survey of some of the promising SDN/NFV based architectures for the Radio Access Network (RAN) and highlight how these architectures can be utilized to support features like network virtualization and slicing. We also underscore the gaps which need to be addressed by these proposals to be able to support the 5G network capabilities and list a few considerations for slicing the 5G RAN. Finally, we propose VirtRAN, a recursive SDN/NFV based architectural framework for RANs, which addresses some of these gaps and can be used to support features like network slicing and user mobility management in 5G networks in an efficient manner.

Index Terms

Multi-RAT, 5G, network slicing, software defined wireless networks, virtualization, recursive architecture

I. Introduction

The upcoming Fifth Generation (5G) mobile network has been designed to support services with diverse latency and throughput requirements. At present, four main use cases viz., enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communications (URLLC), massive Machine Type Communications (mMTC) and Vehicle to Everything (V2X) are being envisaged with flexibility to add other services in the future. However, the cost of network upgrades from present-day Third Generation (3G) and Fourth Generation (4G) Long Term Evolution (LTE) mobile networks to 5G mobile networks is very high especially in the Radio Access Network (RAN) [1]. This is in contrast to the decrease in Average Revenue Per User (ARPU), especially in emerging markets like India [2]. Therefore, to provide an evolutionary path from 3G/4G to 5G networks without incurring a considerable expenditure in network upgrades, service providers are looking at a two-fold approach viz. exploring network architectures with increased flexibility to support new applications/services that may emerge in future and secondly, supplementing cellular network capacity through deployment of low cost solutions in unlicensed spectrum, such as Wireless Local Area Networks (WLANs) leading to a multiple-Radio Access Technology (multi-RAT) heterogeneous network.

In order to create flexible networks that provide assured Quality of Experience (QoE) to users for different types of services, the telecom industry is increasingly turning to softwarized technologies such as Software Defined Networking (SDN) and Network Function Virtualization (NFV). SDN is a networking paradigm that originated through experimentation in Internet Protocol (IP) based networks and has since then been extensively used in data centers [3]. As defined in [4], SDN is -*'A programmable networks approach that supports the separation of control and forwarding planes via standardized interfaces'*. This separation of functionality through a standardized interface enables each plane to be scaled independently. It can also lead to logical centralization of network control at an entity called the SDN controller, having a global view of network resources. SDN based network architecture also defines an independent application plane on top of the control plane [4]. These two planes are connected through a standard interface, typically a Representational State Transfer (REST) based Application Programming Interface (API). Due to the presence of standardized interfaces between the three planes (application, control and forwarding planes), development and deployment of new applications in the network can be achieved with ease. This is unlike traditional networks where all three planes may be tightly coupled and deploying newer applications may not be straightforward [5].

SDN is often used in conjunction with NFV for providing increased network flexibility. NFV is defined by European Telecommunications Standards Institute (ETSI) specifications [6] as - *'the principle of separating network functions from the hardware they run on by using virtual hardware abstraction'*. NFV is used to decouple/virtualize network functions that run on specialized hardware into virtual network functions/software that could be executed on general purpose hardware. While SDN provides interface separation across control and forwarding planes, NFV helps in their dynamic instantiation and scaling. NFV provides the ability to virtualize hardware resources such as storage, compute, network, devices, etc., and enables flexible allocation of resources to different network functions.

Both SDN and NFV enable the creation of logical (virtual) networks, also known as network slices, over a shared infrastructure [7]. Each network slice can be used to support a different service without affecting other slices. This allows service providers to deploy newer services over a common network infrastructure without affecting existing services. A user can access multiple slices concurrently to avail different services at the same time.

Although both SDN and NFV have successfully been used in wired networks to bring advantages such as flexibility (e.g., creation of virtual networks), independent evolution of functional planes (control and data planes) and efficiency (improved throughput, better QoE, etc.), their usage in wireless networks, especially in the Radio Access Network (RAN), presents a different set of challenges. Many of these challenges stem from factors such as - broadcast nature of the wireless medium, need for supporting user mobility, complexity of radio access protocols etc.

In this paper, we present an evolutionary overview of SDN/NFV based architectures for RAN. We discuss some of their strengths and also highlight the issues primarily from the perspective of virtualization and network slicing. As most of the existing solutions for network virtualization and slicing are tailored for a specific network architecture [8] and do not provide insights into the important factors for slicing other architecture types, our paper tries to address this issue. We also propose VirtRAN, a generalized SDN/NFV based framework for slicing multi-RAT RAN.

In summary, the main contributions of this paper are as follows:

- The paper provides an overview and analysis of SDN/NFV based RAN architectures and their application to RAN slicing.
- It provides details of ongoing efforts in a few key Standards Development Organizations (SDOs) towards the development of SDN based multi-RAT architectures and the provisions

for RAN slicing within those architectures.
- It presents the limitations of existing architectures with respect to RAN slicing and proposes the mechanisms for slicing multi-RAT RAN architectures in a generic manner.
- It proposes a novel SDN/NFV based framework for wireless networks with a focus on RAN slicing.

The paper is structured as follows. Section II provides details of existing works on Software Defined-RAN (SD-RAN) architectures. It is followed by Section III, which provides a brief overview of the 3GPP 5G network and its adoption of SDN both within the core and RAN. The section also discusses the adoption of network slicing in 3GPP. In the subsequent section, we provide an introduction to Software defined wireless networks and slicing in other SDOs. Section V discusses some of the open issues in slicing multi-RAT RANs while summarizing some guidelines for better ways to slice a RAN. This is followed by Section VI, which describes a possible architectural solution for slicing various network deployment models. An implementation of a rate-limiting mechanism for VirtRAN using ns-3 is described in Section VII. We present a few advantages of the proposed architecture in Section VIII followed by conclusions in Section IX.

II. Existing works on SD-RAN and Network Slicing

In this section, we review some of the existing literature that has had a significant impact on software defined RANs and network slicing. As discussed previously, the next-generation network consists of heterogeneous RATs such as LTE, New Radio (NR), and WLAN. Therefore, we discuss works describing the individual RATs as well as multi-RAT architectures.

A. *Software defined WLAN*

Odin [9] is an implementation of SDN controlled WLAN and is amongst the first works to illustrate the working of a user-centric WLAN architecture. The Odin architecture consists of a controller controlling Access Points (APs) within WLAN. AP control is achieved with the help of two agents running on the APs i.e., Odin agent for radio control and OpenFlow agent for flow control. The authors also introduce the concept of a Lightweight Virtual Access Point (LVAP), which is an abstraction of the WLAN AP for a given user. LVAPs are installed on APs and are assigned to a given user by the Odin controller. As a result, it appears that each user has been assigned a dedicated AP. An LVAP is responsible for all communication with a single User Equipment (UE). LVAPs are migrated across APs when a user moves and through this mechanism, handovers across APs are transparent to a given UE. The Odin framework also demonstrates a method for slicing WLAN. Within this framework, a network slice consists of a set of WLAN APs, clients with their associated LVAPs and Service Set IDentifiers (SSIDs) along with the associated network application.

Despite its elegance, the solution has certain shortcomings. The solution may not be scalable for a large number of UEs, as beacons are sent over multiple unicast channels to individual UEs by the corresponding LVAPs. Also, the proposed scheme may not work when the target AP is on a different channel than the source AP as mobility is no longer transparent to the UE. Moreover, this scheme may not be easily adapted for use over LTE and 5G-NR due to the changes in protocol required.

OpenRoads [1] is an early work that demonstrated SDN based control and management of multi-RAT networks, e.g. WLAN and Worldwide interoperability for Microwave Access (WiMAX). In this solution, the authors propose the usage of OpenFlow for configuring routes

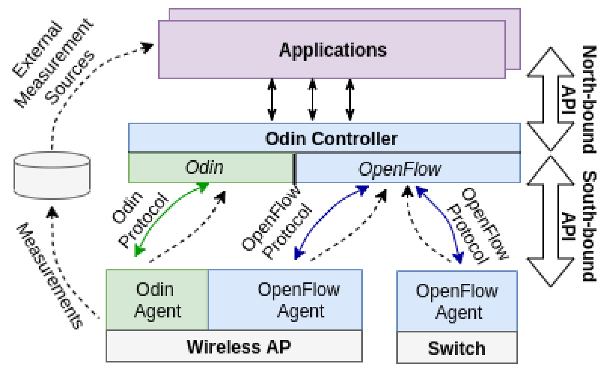


Fig. 1: System architecture for Odin (Courtesy [9]).

within a network and the use of Simple Network Management Protocol (SNMP) [10] to configure radio-related parameters on the APs. This architecture treats WLAN AP as an OpenFlow switch. OpenRoads architecture comprises three layers, i.e. flow layer consisting of OpenFlow tables and SNMP, slicing layer consisting of a virtualization solution known as Flowvisor [11] for slicing the network and controller layer consisting of a controller for centralized network control. The Flowvisor is placed in between the APs and SDN Controllers and divides the flow-space manifested by individual APs into smaller sub-spaces and maps these individual sub-spaces to separate network slices. Individual network slices may be controlled by different SDN Controllers. OpenRoads architecture is quite similar to the one proposed by Open Networking Foundation (ONF) discussed in Section IV. For example, Flowvisor can be viewed as a lower level SDN Controller as specified in the ONF architecture.

Since IEEE 802.11 MAC layer has many similarities to Ethernet MAC, it is possible to view WLAN APs (with 802.11 MAC) as Ethernet switches and use OpenFlow protocol to control them. However, there are inherent limitations in using flow level abstraction as the interface between control and data planes in wireless networks for a few reasons. For example, if we use this interface, the allocation of underlying radio resources, e.g., bandwidth to each of the network slices (flow space) is completely hidden from the SDN controller. Instead, it becomes the responsibility of the data plane entities (APs), thereby defeating the purpose of having an SDN based network architecture. Also, as Flowvisor is responsible for creating slices over the flow-space manifested by APs, APs themselves are unaware of the network slices. Further, due to time and user-specific variation in radio channels, the allocation of radio resources to different slices may vary over time. Due to APs' unawareness of the network slices, they are unable to maintain slice specific separation over radio resources. As far as other mobile technologies, such as LTE and 5G NR are concerned, the applicability of OpenFlow as the interface between the RAN control plane and data plane functions is even more limited as their radio protocol structure is more complex in comparison to IEEE 802.11 WLANs. Besides, they also use abstractions such as radio bearers, which may need to be manipulated by the Controller.

Lasagna [12] is another WLAN slicing solution defined over the 5G-EmPOWER [8] SDN based multi-RAT testbed designed to control LTE and WLAN. The authors implement slicing by using a programmable hypervisor on top of the Linux WLAN stack. They also extend OpenFlow match-action rule for use in WLAN environments and introduce a new abstraction known as 'traffic rule' for mapping a given portion of the flow space with a scheduler. The traffic rule identified by a tuple comprising SSID for identifying destination WLAN and Differentiated

Services Code Point (DSCP) for indicating priority for an IP packet.

B. Software defined LTE networks

SoftRAN [13] is an early work for SDN based architecture for cellular RAN. It is similar to a centralized Self Organizing Network (SON)/ Radio Resource Management (RRM) solution proposed by 3GPP. In this work, the authors propose a hierarchical SDN controller for dense cellular deployments. The control functionality of a base station, which may have an impact on its neighboring cells such as handover decisions, transmit power control for mitigating interference, UE uplink resource block allocations, etc., are abstracted out into a global controller. The control functionality for localized decisions of a physical base station is then abstracted into a local controller. The underlying network resources are abstracted into a three-dimensional resource grid of base station index, time, and frequency slots.

A solution for slicing SoftRAN was proposed in Radiovisor [14]. Radiovisor highlights the fact that for wireless networks, interference is an additional factor for slice creation and management, unlike those of wired networks. Hence, spectrum resources allocated for each slice must be isolated and not interfere with one another. Radiovisor also supports the inclusion of a per-slice controller and application(s) and deployment of layered configuration e.g., scheduling for MAC, physical layer configuration, etc. for a specific slice flexibly and independently. However, the procedure for slicing control plane resources and ensuring isolation is unclear from the information provided.

FlexRAN [15] is a proposal that introduced and implemented the idea of software defined RAN for cellular networks. Although FlexRAN has been designed and implemented for LTE networks, the authors state that it is extensible for future RATs and also describe some of the necessary steps for the same. As illustrated in Figure 2, it is a hierarchical architecture with a centralized master controller and a FlexRAN agent (local controller) deployed at each LTE base station i.e., the eNB. The control functionality within Radio Resource Control (RRC), Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), and MAC are transposed into the master controller. The master controller can perform scheduling and radio resource scheduling decisions centrally for eNBs under its control. The proposed architecture also provides the flexibility to use FlexRAN in bandwidth constrained environments by introducing control modules known as Virtual Subsystem Functions (VSFs) within the FlexRAN agent where both scheduling policies and resource configurations can be provided and updated. This allows for localized operation at eNBs when necessary.

Cloud RAN [16] is another popular solution that was initially proposed by International Business Machines Corporation (IBM). It involves the centralization of baseband processing of a base station in a centralized datacenter. As the initial proposal specifying separation of the baseband and radio terminals requires a high-speed link ($>$ 1.5 times the bandwidth) which is more bandwidth intensive in comparison to the back haul [16], various other configurations with different stack separations have also been introduced in this work.

C. Software defined multi-RAT networks

5G-EmPOWER [8] is one of the first works that has implemented SDN based multi-RAT controller. The solution provides a framework with an SDN controller as a network operating system to control and manage LTE and WLAN with the help of a unified controller. The proposal defines a new management protocol known as OpenEmpower and an Operating System (OS) known as 5G-EmPOWER. The architecture transposes management functionalities from the RAN

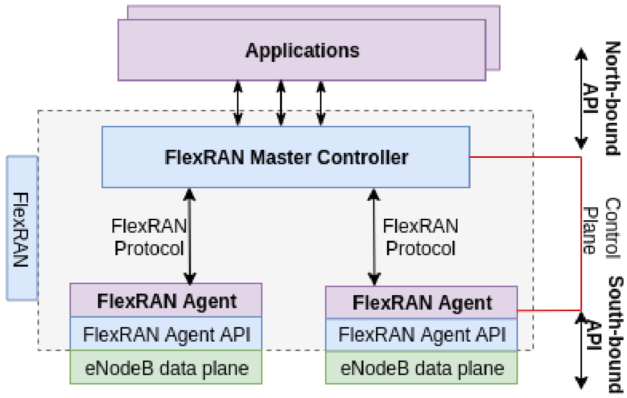


Fig. 2: System architecture for FlexRAN (Courtesy [15]).

nodes and moves them to the management plane running over the 5G-EmPOWER operating system. The 5G-EmPOWER operating system behaves as a controller, and a 5G-EmPOWER agent is placed on each RAN node so that it can be configured by the OS. The 5G-EmPOWER OS is responsible for functions such as allocating data plane resources for users, providing isolation between users, providing a RAT-agnostic view of resources to users by abstracting network resource details etc.

The 5G-EmPOWER framework also demonstrates RAN slicing for LTE network. The proposed slicing mechanism places a hypervisor above the physical layer. The hypervisor performs the abstraction of Physical Resource Blocks (PRBs) into virtual PRBs, which can then be grouped into virtual PRB groups for use. A slice resource manager placed at the MAC layer above the hypervisor is used for managing the slice. Multiple slices with independent schedulers can be created at the MAC. The virtual PRB groups created with the help of the hypervisor are then mapped to be allocated to be used by slice specific schedulers for performing slicing. However, the authors do not provide details on how slicing could be performed over WLAN. Similar to SoftRAN, the 5G-EmPOWER framework has been developed with a view to support centralized SON server functionality over Multi-RAT RAN.

III. SDN and NFV in the 3GPP 5G Network

The 3GPP 5G network architecture incorporates both SDN and NFV elements in its architecture. A few SDN principles have been introduced into the core network in Release-14 of LTE through control and user plane separation specification [17]. Within this specification, certain core network elements such as Serving Gateway (SGW) and Packet Gateway (PGW) are split into control (SGW-c, PGW-c) and data plane elements (SGW-u, PGW-u) according to their functionalities. The 3GPP 5G architecture further refines these ideas by introducing a service-based architecture where all the elements within the network are made up of network functions. Network functions are elements that have well-defined functions and interfaces and can be implemented either in hardware or software [18]. As illustrated in Figure 3, the core network comprises SDN based function separation wherein the User Plane Function (UPF) has data plane functionality and all other core network functions such as Access and Mobility Function (AMF), etc., have control plane functionality. In the RAN, 5G base stations known as gNB are constituted by gNB-Centralized Unit (gNB-CU) and gNB-Distributed Units (gNB-DUs). gNB-DUs have pure data plane functionalities, and gNB-CUs have both control and data plane functionalities.

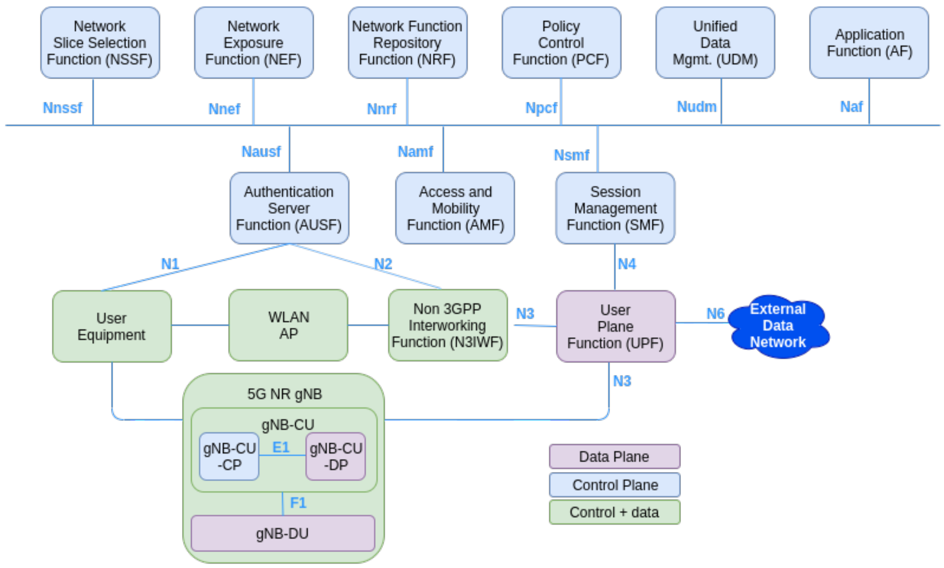


Fig. 3: 3GPP 5G system architecture (Courtesy [18]).

A gNB-CU can be further subdivided into a function having control plane functionality viz., the gNB-CU Control Plane (gNB-CU-CP), and gNB-CU Data Plane (gNB-CU-DP) having only data plane functionality. The gNB-CU-CP can configure the gNB-CU-DP with the help of F1 interface [19]. F1 interface is used for the separation of control plane and data plane functions of the gNB-CU and gNB-DU. It can be used to configure radio bearers and is further split into the control and user plane parts viz., F1-C and F1-U, respectively. The gNB-CU and the gNB-DU communicate with the help of E1 interface. F1 Application Protocol (F1AP) [20] is used over F1 interface and E1 Application Protocol (E1AP) [21] is used for communication over E1 interface. Although at first glance, F1AP seems like OpenFlow as it is used between entities with control and data plane functionality split, it is a specific protocol developed for use in 5G NR to carry messages for radio bearer configuration and transporting data. Similarly, in non-3GPP access networks, Non 3GPP Interworking Function (N3IWF) has a mix of both control and data plane functionalities.

The 3GPP specifications [18], [22] perform end-to-end slice selection in the following manner. During the initial attach, the UE can indicate the list of slice identifiers that it supports using Network Slice Selection Assistance Information (NSSAI) and request for service on specific slices. The RAN uses NSSAI provided by the UE during 'RRC Connection Establishment' procedure to select a suitable AMF that supports these NSSAIs. If this information is not available from the UE, a default AMF is chosen by the RAN. If the first contacted AMF does not support all the slices requested by the UE, a change of AMF may be initiated. At present, a UE may simultaneously access 8 slices. The AMF is common across all slices for a given UE whereas other functions such as SMF and UPF may differ. However, the core network is responsible for ultimately deciding the allowed NSSAIs. Therefore, once a UE successfully registers with an AMF, the core network informs the RAN of the allowed NSSAIs for the given access type. Although mechanisms for core network slicing have been defined by 3GPP, the choice of network functions for a specific slice within the RAN is at present deemed to be implementation dependent [22].

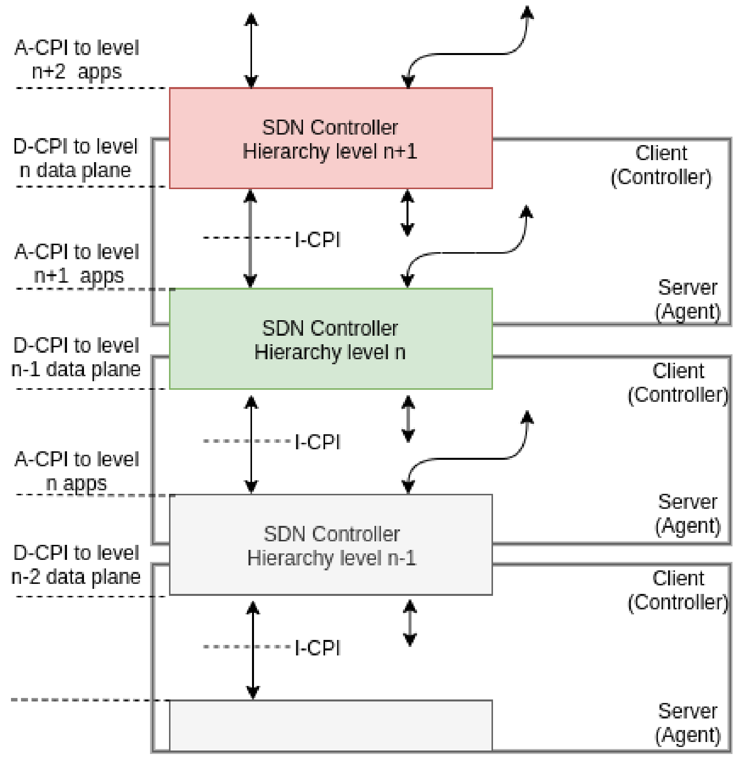


Fig. 4: ONF recursive architecture (courtesy [23]).

IV. Other frameworks for SDN and Network Slicing

Architectural frameworks for SDN have also been defined on similar lines by both Open Networking Foundation (ONF) [23] and Internet Engineering Task Force (IETF) [4].

ONF framework enables SDN controllers to be placed in a recursive or hierarchical fashion for better scalability, as illustrated in Figure 4. Within a recursive framework, the higher-level controller say at level $n+1$ appears to the lower level controller n as an application. Similarly, the controller at level $n-1$ appears as data plane to controller at level n. Recursion allows for the creation of applications that provide finer-grained services by combining multiple applications. Recursive placement may also be used to provide different security levels within the network. The ONF specification TR-526 describes the concept of network slicing [24]. It highlights the fact that features such as resource virtualization, recursion, etc., provided by SDN based architectures make them ideal for the implementation of network slicing vis-a-vis traditional networks.

The division of the flow-space into smaller sub-spaces using OpenFlow can be utilized in a recursive network architecture, where SDN controllers are organized into a recursive hierarchy with a lower-level controller (the one closer to the data plane) responsible for subdividing the flow-space into sub-spaces and mapping these individual sub-spaces to independent virtual networks. Each one of these virtual networks (sub-spaces) may further be exposed to separate higher-level controllers for management and control purposes. The virtual network controllers can manipulate the corresponding virtual networks through OpenFlow protocol.

A widely used element to aid network slicing is the hypervisor. A hypervisor is used for abstracting the physical network to create one or more isolated virtual networks that can be controlled individually. The creation of multiple networks can be achieved by placing the hypervisor

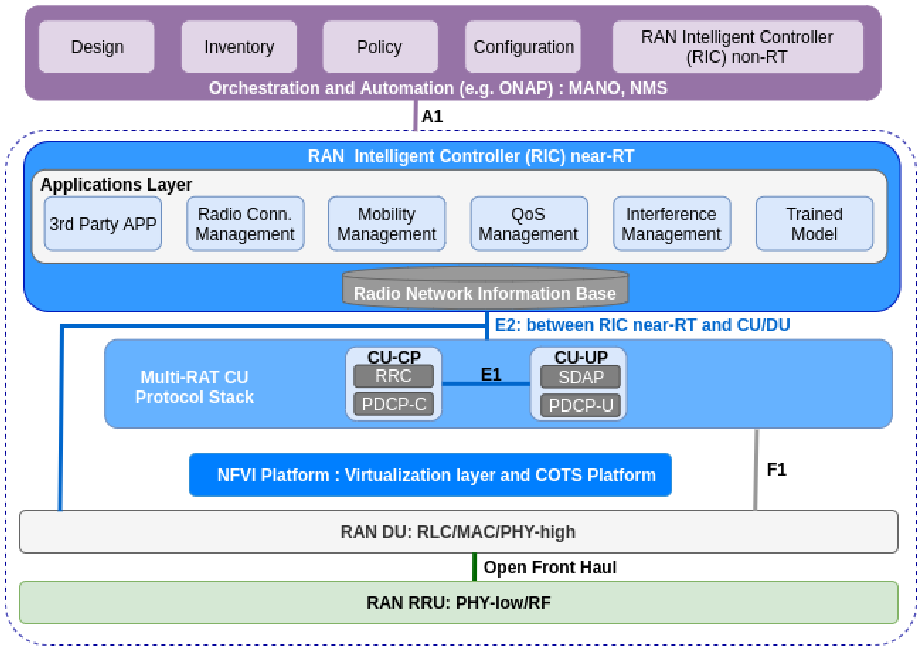


Fig. 5: O-RAN architecture (courtesy [26]).

between the data plane and the controller plane. For example, in Figure 4, the hypervisor for SDN controller at level n can be placed at the place of SDN controller at level $n-1$. Hypervisors are used for reducing the complexity of network control by providing a simpler view of the network in terms of topology, network resources, etc. A detailed analysis of the various types for SDN networks is provided in [25].

Other than the frameworks discussed above, several standardization activities for achieving control and management of multi-RAT networks are underway. O-RAN is a standard that is presently under development under the newly formed O-RAN alliance [26], a consortium of cellular network operators. The objective of O-RAN is to develop an SDN based smart RAN with open interfaces for enabling vendor inter-operability and usage of artificial intelligence/ machine learning algorithms for optimized network decisions. O-RAN APIs and interfaces are defined using 3GPP specifications as their basis. This standard promotes the usage of open-source software and off-the-shelf hardware for reducing CApital Expenditures (CAPEX).

The system architecture of O-RAN is illustrated in Figure 5. The radio interface control functions in O-RAN are decoupled as non-Real Time (RT) ($> 1s$) and near-RT ($< 1s$) based on the time scale of operation. The non-RT Radio Interface Controller (RIC) is responsible for longer time-scale decisions such as policy management, configuration, training of learning models from the collected data, etc. On the other hand, the near-RT RIC interfaces with the non-RT RIC through A1 interface and provides RRM related functionality such as mobility management, Quality of Service (QoS) management, etc. It also enables third-party applications to be easily incorporated into the network and maintains a near-RT network state by gathering data from the layers below through the E2 interface. O-RAN supports 4G LTE and 5G NR RATs at present. As within the 3GPP 5G specs [27], the radio protocol stack has been split into CU and DUs. The interfaces defined by 3GPP, such as E1 (between control and data plane) and F1 (between CU and DU), are being extended for use within the O-RAN standard. The first release of O-RAN code-named 'Amber' has been released in November 2019. O-RAN is built as an

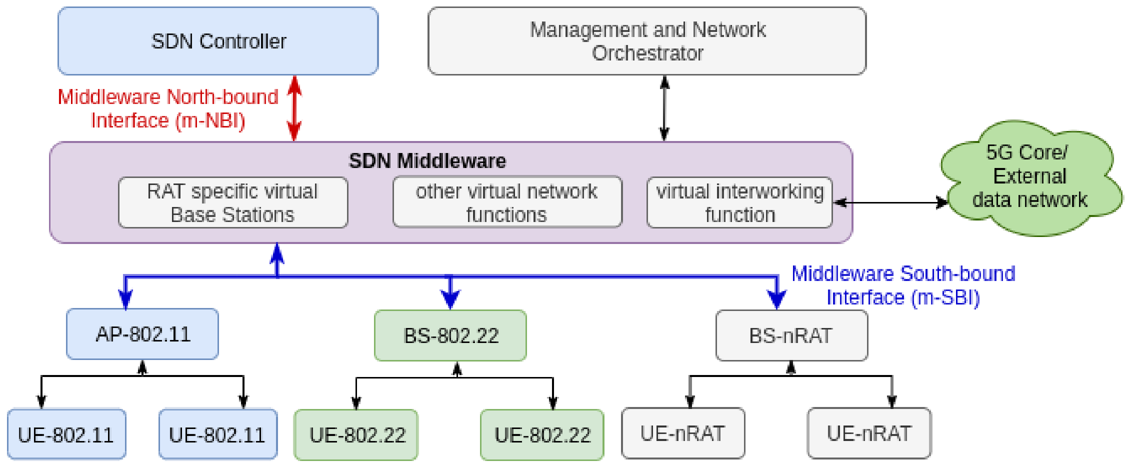


Fig. 6: SDN based Middleware Architecture defined in P1930.1.

extension to 3GPP and hence does not provide any specific guidelines for slicing the RAN. It is intended that the mechanisms defined by 3GPP would be used as is unless explicitly mentioned within O-RAN specifications [26]. As a result, it is inferred that slicing within O-RAN is also implementation dependent.

Another on-going standardization activity aimed at improving vendor interoperability and enabling control of IEEE multi-RAT networks is P1930.1 [28]. The objective of this standard is to define an SDN based middleware architecture for control and management of multi-RAT networks such as 802.11 WLAN [29] and 802.22 Wireless Regional Area Networks [30]. A similar standard known as Control and provisioning of wireless access points (CAPWAP) [31] was previously defined in IETF. However, the specifications for a particular RAT (known as bindings) are only available for WLAN [32] at present. Also, the standard is not SDN compliant and has not been widely adopted by operators.

The system architecture of a P1930.1 compliant network is illustrated in Figure 6. The standard introduces an SDN based middleware for multi-RAT networks. The middleware has two sets of interfaces. The interface between the SDN Controller and the middleware, i.e. the middleware-NBI uses two protocols - OpenFlow and NETCONF. NETCONF is used for configuring radio parameters of AP and base stations whereas OpenFlow is used to set up flows in the wired portion of the network. On the south bound interface, existing protocols such as CAPWAP or SNMP can be used. This allows P1930.1 middleware to be easily integrated into present-day WLAN networks while enabling vendor interoperability.

The middleware introduced in P1930.1 segregates control and data plane functionality within the network. As a result of this, the resulting network can be controlled using an SDN controller. The middleware effectively acts as an access controller [31] used in today's networks for controlling the APs. The middleware is controlled by an SDN controller and can be orchestrated with the help of an orchestrator.

Although the standard is being defined for WLAN and WRANs, this standard has been envisioned to be extensible to other RATs in future. Moreover, the standard also provides an ability to integrate non-3GPP networks with 3GPP 5G core network. As the standard does not introduce any changes in the UE protocol stack, present-day UEs can be used within this network.

P1930.1 makes provisions for slicing the WLAN by virtualizing the APs and the 802.22 base stations within the middleware as virtual Base Stations (vBS). These vBSs are orchestrated with the help of an existing orchestrator and provide an abstract view of the network resources to the SDN controller. The SDN controller can then use this view to partition these resources for slicing the network.

V. Open issues in slicing multi-RAT networks

In the previous section, we have reviewed some of the existing works on network slicing in multi-RAT networks. In this section, we identify a few gaps that exist in the literature used for controlling SDN based multi-RAT networks. We also analyze the improvements required within them to support network slicing.

As mentioned in an earlier section, OpenFlow in its present form is not suitable for usage in software defined wireless network environment. This is due to the fact that the concept of flows may not be able to capture some of the distinctive characteristics of the wireless networks. With a flow level interface, the allocation of the underlying resources, e.g., bandwidth, compute or storage to each of the network slices (individual flow spaces) is not visible to the SDN controller and the controller is unable to allocate these resources directly to the network slices. Even though the controller configures flows on the data plane nodes, the responsibility of actual resource allocation to individual flows is with the data plane. While this issue does not occur in wired networks as there is minimal variation in the link quality across time and users, this may not be the case in wireless networks. Due to time and user-specific variation in radio channels, the number of underlying resources allocated to individual flows and therefore, to the network slices (flow ensemble) varies over time, which may defeat the concept of slice separation.

To illustrate this problem, let us consider an SDN based LTE network where two users are accessing a Youtube video over a smartphone using LTE. Let us assume that one of the users (user 1) is in the centre of an LTE cell, near the base station while the other user (user 2) is at the cell edge, far from the base station. Let us assume that an SDN controller configures the network into two slices and configures a policy for providing equal resources to them. Let us assume that user 1 is accessing slice 1 and user 2 access slice 2. However, the user at the cell edge (user 2) is likely to experience a poor radio link as compared to the user at the cell centre (user 1). Even though the policy was to provide equal resources to the flows, due to adverse channel conditions of user 2, the base station would allocate a larger amount of resources to user 2 for compensating for the channel conditions. Further, resource allocation may vary over time, especially if users are mobile. If the resources present in slice 1 are inadequate for compensating for the channel conditions, some of the resources may be drawn from slice 2, thus affecting slice isolation. As described above, the network policy proves ineffective and demonstrates that granular control may not be achieved only through flow-configuration in Wireless networks.

Another important factor for ensuring optimal system performance in wireless networks is interference management. In order to achieve this, operators not only have to optimize throughput by configuring flow paths but also manage radio parameters such as transmit power etc. OpenFlow does not have provisions for configuring radio related parameters and may need to be extended or supplemented by another protocol such as NETCONF, SNMP, etc. Although some work has been done to address this issue for WLANs [33], this still remains a challenge in cellular networks such as LTE and 5G-NR.

Works such as Radiovisor designed over SoftRAN architecture [13] have proposed mechanisms for slicing LTE networks. However, in Radiovisor, the resources have been statically split right down to the physical layer. Moreover, the work does not provide enough information on splitting the control signaling (e.g., broadcast information signaling) for achieving slice isolation. Unlike OpenFlow, which can be introduced in an SDN based network without impacting the existing protocol stack, this is not possible with SoftRAN architecture. Consequently, the resulting SDN implementation may not be standard compliant and thus not interoperable.

Additionally, with a recursive architecture as proposed in ONF [23] and OpenRoads [1], data plane may not have an awareness of network slices, as the division of flow space into different network slices is visible to the Controller only. Therefore it may be challenging to adhere to any resource separation at the slice level in such a scenario. Apart from this, a few other issues regarding network slicing are also presented in [34].

Based on the discussion so far, we identify the following characteristics for a better design of slicing architectures.

- The architecture should support the concept of abstract network resources, which can be used for virtualization and network slicing.
- The abstract network resources should not be wholly unconnected from the underlying resources that are being represented and allocated. For example, the concepts of traffic flow, as defined in OpenFlow and used in OpenRoads, is a simple and beautiful abstract resource but it may not be capable of accurately representing the radio resources in wireless access networks.
- The architecture should provide a flexible virtualization scheme so that different mechanisms (simpler to more complex) can be used for virtualization depending on the use case/requirements to be supported.
- The architecture should support the virtualization of resources at multiple levels as resources can be defined and grouped at different levels. For example, the LTE physical layer takes a frequency band and represents it as PRBs to higher layers. Similarly, the MAC layer may take the PRBs and represent them as virtual resource blocks (VRBs). The architecture should have the flexibility to work with such different types of abstractions.
- It should support an SDN based architecture with separate control and data plane functions with an open interface between the two.
- The architecture should support a clean separation between the control plane and the data plane and allow for data plane virtualization at multiple levels without the presence of the control plane functions in the data path. This is different from the SDN based architecture as proposed by ONF.

VI. Proposed framework for SDN based multi-RAT network

With the help of guidelines provided in Section V, we propose VirtRAN, an SDN framework that could be used for slicing multi-RAT networks. This framework enables us to virtualize the network at multiple layers in order for better slicing the network.

A. VirtRAN: SDN based framework for recursive multi-RAT network slice deployment.

As illustrated in Figure 7, VirtRAN has a well-defined separation of control and data planes. The data plane is further subdivided into multiple sub-planes, wherein each sub-plane performs a part of the data plane functionality and utilizes a set of resources. There may be an optional "Virtualization Layer (VL)" over each one of these sub-planes. The VL at a given level creates an abstract view of the underlying resources and provides these resources/ groups of resources to the layer above for use. It is also responsible for ensuring isolation across resource groups. It not only virtualizes radio resources but may also virtualize compute and storage resources for higher sub-planes. As illustrated in Figure 7, each network slice (resource group) at a sub-plane level may be controlled/managed by a separate slice specific SDN Controller.

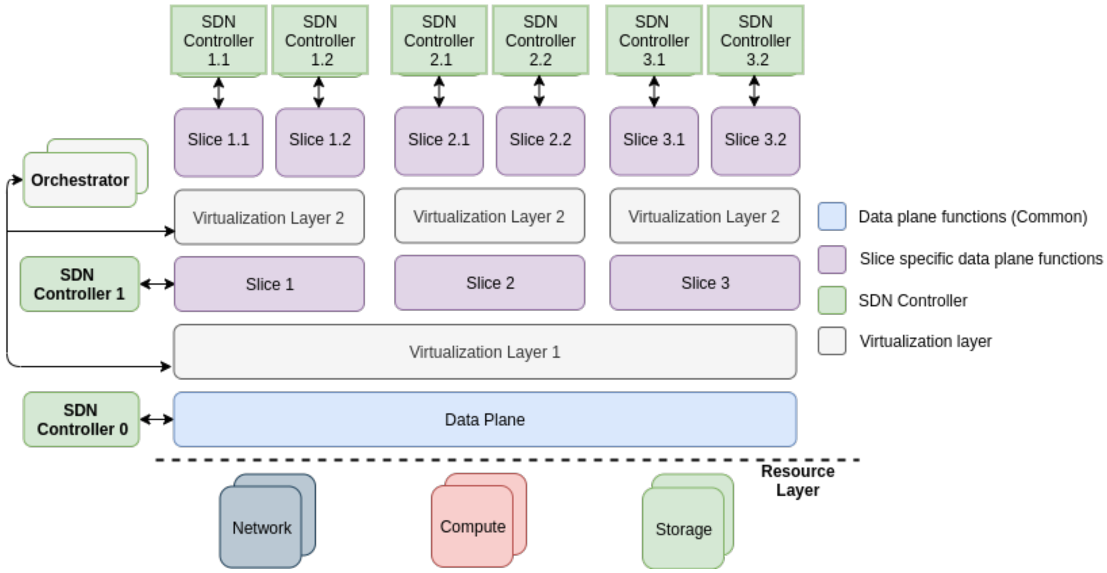


Fig. 7: Proposed recursive architecture for slicing.

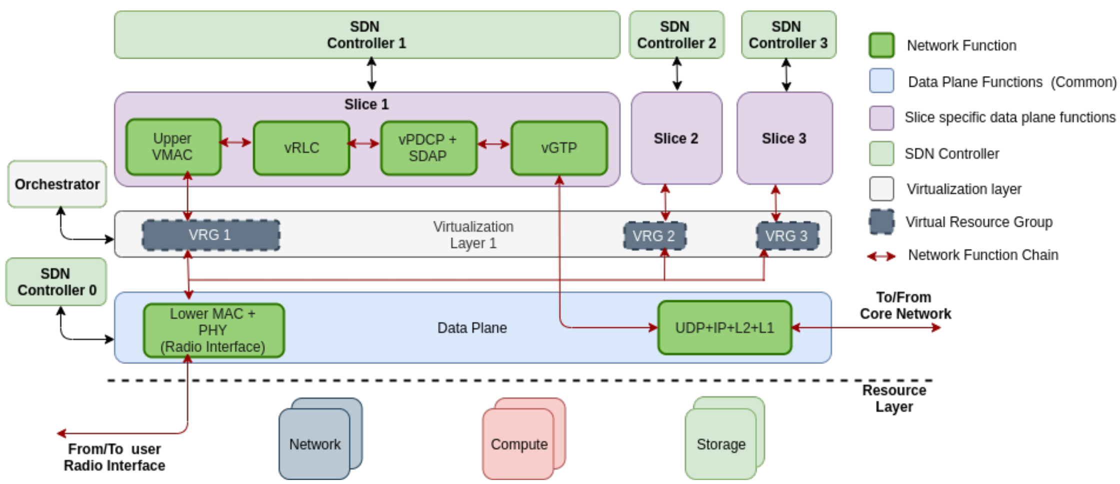


Fig. 8: Data Path illustration for 5G NR in the proposed recursive architecture.

VL acts under the control of an orchestration entity known as the orchestrator. The role of the orchestration entity may also be played by the SDN controller responsible for the control of the underlying sub-plane (just below the VL). In this case, the SDN controller manages/controls both the virtualization layer and the underlying sub-plane.

We discuss further details of the proposed architecture with the help of an example of a network consisting of LTE/ 5G NR technology, as illustrated in Figure 8. In this case, a possible division of the radio interface in sub-planes is to group the radio protocol layers and put them into a separate sub-planes, e.g., Physical Layer and the MAC Layer could be a part of single sub-plane whereas RLC, PDCP (and Service Data Adaptation Protocol (SDAP) in case of 5G) layers may be a part of another sub-plane.

The sub-plane containing the RLC/PDCP layers may also contain other layers such as General Packet Radio Service (GPRS) Tunneling Protocol - User Plane (GTP-U), etc., which are not part

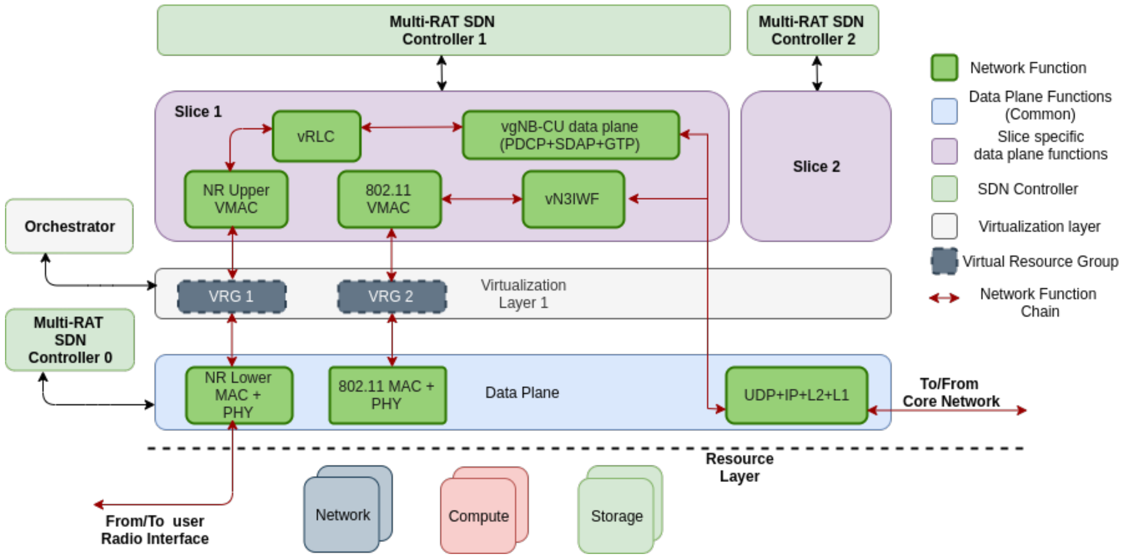


Fig. 9: Example of virtualization of the proposed architecture for 5G NR and WLAN.

of the radio interface but the core network interface. In this case, a VL may also be placed over the PHY and MAC layers and can be responsible for virtualizing the underlying physical radio resources into subsets of (virtual) radio resources.

The VL can also divide the virtual resources into multiple subgroups and allocate each of the resource subgroups to a virtual network or network slice. There may be multiple network slices comprising of individual RLC and PDCP layers. Here a slice specific MAC layer functionality would also be required in the higher sub-plane to allocate virtual radio resources to slice specific users. A slice may also contain other layers, such as SDAP and GTP-U. Each of the slices utilizes one of the resource subgroups.

In Figure 9, there are slice specific virtual MAC, RLC, PDCP, SDAP and GTP functions. The slices shown may correspond to three of the four 5G use cases, e.g., eMBB, mMTC, and URLLC. The hypervisor over the MAC layer divides the underlying radio resources into three different groups. Each group is mapped to a corresponding virtual resource group and presented as a slice. The slice specific virtual MAC function is mapped to one of these resource groups by the virtualization layer. However, there is no limitation on the number of slices and virtual resource groups, and the VL may be configured to support any number of slices/resource groups. It is possible for the proposed architecture to support a unified Multi-RAT RAN. The VL running over the LTE MAC and the IEEE 802.11 MAC layer may be able to unify the underlying RAT specific physical radio resources through an abstract/virtual resource view created for the higher sub-planes. This would enable the higher sub-plane data functions, e.g., RLC, PDCP layers, to utilize either or both of the underlying LTE or 802.11 MAC layers seamlessly.

The division of data plane into multiple sub-planes, virtualizing the resources at multiple levels (at each sub-plane level) and putting each sub-plane under sub-plane specific SDN controllers enables granular control over the resources being used in the network as opposed to the architectures such as the one proposed by ONF, wherein there may be a single level of abstraction used over the data plane resources.

As shown in Figure 7, it is a recursive architecture wherein there may be a VL over each of the slice specific sub-planes. The VL over each slice may further divide the underlying slice

TABLE 1: Simulation parameters.

Parameter	Value
Path loss	$128.1 + 37.6 \log(R)$, R in kms
Tx power for LTE dBS	46 dBm
Bandwidth of the LTE dBS	10 MHz
Tx power for UE	23 dBm
Antenna Type LTE	Isotropic Antenna

specific resources into smaller resource groups and export them for manipulation to another higher level slice specific functions. Although most of the discussions here have been in the context of RAN, it is evident that the concepts explored here are generic and can be applied to the core network as well.

VII. Simulation of VirtRAN

In this section, we provide simulation results for VirtRAN compatible LTE Radio Access Network. The simulation is carried out for LTE-RAN using ns-3 [35] LENA module. Note that we use LTE for simulations, as to the best of our knowledge, a full stack simulator for 5GNR is not yet available.

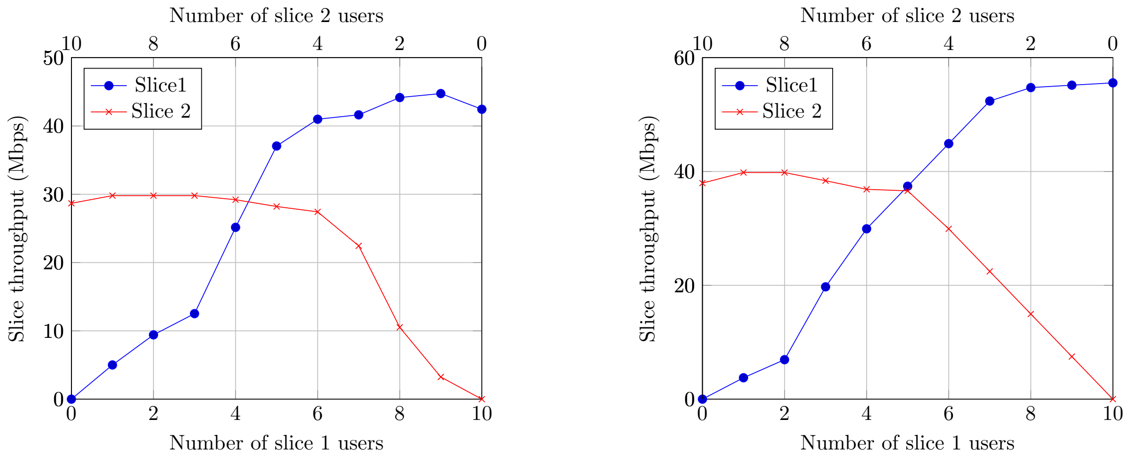
The VirtRAN concepts, as proposed in the previous section, are illustrated through the creation of network slices in LTE RAN. To perform the simulations, we virtualize the resources at the LTE PDCP layer and divide them into multiple network slices. For virtualization (at the PDCP layer), a resource is defined in terms of achievable data rate. We consider network slicing for downlink data traffic in an LTE base station (eNB). Every network slice in an LTE base station is limited to a configurable maximum data rate. The rate-limiting mechanism aims to reserve a certain percentage of RAN resources and limit the slice throughput to a pre-defined value. The mechanism provides insights on configuring VirtRAN for specific applications.

The simulation setup consists of a macro eNB (release-8) with a single transmit and receive antenna. Hence, the eNB has a maximum capacity of 75Mbps in each direction. Users are uniformly distributed within a cell radius of 100m and access services over slice1 or slice2. Each user accesses an application at a data rate of 10Mbps. We also assume that the number of users for each slice is different and that every user has full buffer traffic. The rest of the parameters used in the simulations are detailed in Table 1.

We perform the rate-limiting at PDCP to achieve a distribution of resources in the ratio 3 : 2 across the slices. To achieve this, we perform the rate-limiting using two different mechanisms. For both the mechanisms, we classify downlink traffic at the PDCP layer as belonging to either slice 1 or slice 2. Traffic flow statistics are monitored for every slice, and once the threshold for a slice is attained, the remaining slice users are not serviced. As a result, the maximum rate for each slice is limited. As our proposal aims to restrict the maximum data rate for the slice only on the downlink, uplink PDCP traffic should remain unaltered and is therefore isolated.

A. Mechanism 1: resource division based on system capacity

In this mechanism, we enforce a hard limit on the resource division based on system capacity. The limiting throughput is set to 45Mbps for slice 1 and 30Mbps for slice 2. The resulting slice throughput for various user distributions is illustrated in Figure 10a. As expected, we observe that the throughput for both slices does not exceed the limit, even when a particular slice is loaded.



(a) Virtualization based on system capacity.

(b) Virtualization based on offered load

Fig. 10: Virtualization at PDCP layer in VirtRAN.

B. Mechanism 2: resource division based on offered load

In the second mechanism, we divide resources in the prescribed ratio based on the offered load. If we consider that there are 10 users within the cell, the offered load is 100Mbps. Therefore, we set a limiting throughput rate of 60Mbps for slice 1 and 40Mbps for slice 2. As illustrated in Figure 10b, this type of allocation provides higher throughput for a slice, especially when the system is not loaded to its capacity, while ensuring resource division at the PDCP level. This also proves that at lower system loads and better channel conditions, virtualization at higher layers of the protocol stack provides a simpler way to achieve network slicing.

From Figures 10a and 10b, we observe that throughput for both slice 1 and slice 2 is within the required range. Figure 11 also illustrates that load based resource allocation provides better system throughput in comparison to the system capacity based fixed resource allocation. These examples demonstrate that network virtualization and slicing implemented at the PDCP layer can achieve the desired goals. This is especially true when the users associated with each network slice are experiencing similar channel conditions and accessing similar types of services.

However, as shown in Figures 11 and 12, there may be certain limitations to the implementation of network slicing at the PDCP layer. This becomes apparent, especially when the network consists of slices providing services of different traffic priorities, and users associated with a slice providing high priority traffic are consistently experiencing poor radio channel conditions. In such scenarios, virtualization at a lower layer (sub-plane), e.g., at the MAC layer, would be necessary to improve resource sharing.

To illustrate this fact, we increase the number of users within the cell. Suppose that one of the slices (say slice 1) is providing only Guaranteed Bit Rate (GBR) type traffic, and slice 2 is serving non-GBR users. If slice 1 users are in general experiencing poor channel conditions than the slice 2 users, a larger number of radio resources are granted to them to maintain the QoS. As shown in Figure Figure 12, even when an equal number of users are present in the slices, we observe that the throughput of slice 2 is negatively impacted even though resources were pre-allocated at PDCP. In order to alleviate this, isolation and allocation of resources could also be performed at lower layers such as MAC, based on the network conditions. VirtRAN allows for recursive virtualization across layers for achieving the fine-grained RAN control as desired

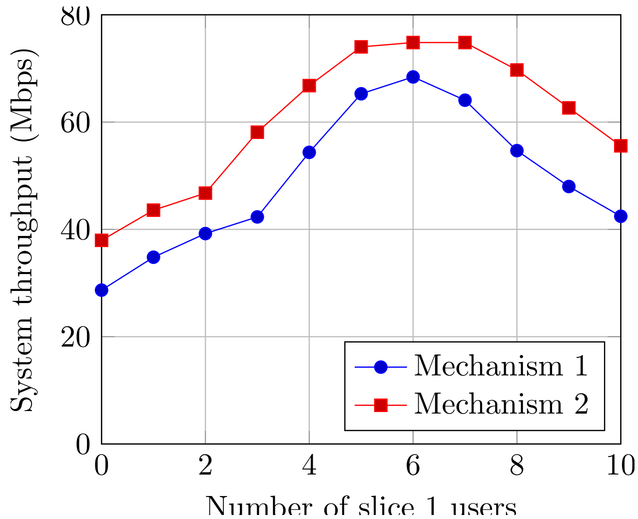


Fig. 11: Comparison of system throughput for both mechanisms.

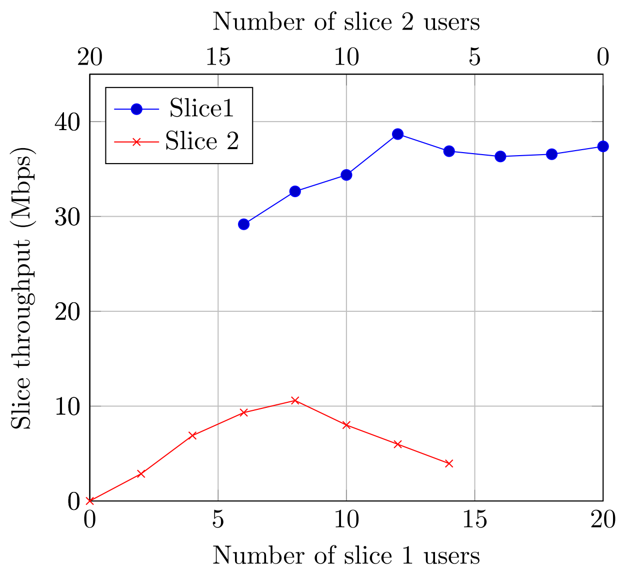


Fig. 12: Illustration of non-isolation of slices.

in this scenario.

VIII. Advantages of the proposed architecture

The proposed architecture has several advantages over existing proposals. Some of them are listed below:

- The framework does not suffer from the flaws proposed by schemes like OpenRoads, which uses a single abstraction mechanism (flow space) for the underlying data plane. The mechanism proposed here can be used to specify data plane abstractions at multiple levels enabling more granular control over resources in the network. The individual data sub-plane abstractions used by the controller to control/manage the individual sub-planes are dependent on the resources being managed by these sub-planes. For example, the sub-plane containing the GTP and PDCP function is controlled through usage of tunnels/bearers/data rates, whereas the sub-plane containing the MAC may be controlled through manipulation of radio resources. Further, placing a VL over the LTE MAC layer and virtualizing the underlying radio resources and allocating them to different slices enables better slice level control over radio resources than an architecture like OpenRoads, which uses the abstraction of flow spaces to manage the radio resources also. Due to better granular control over resources at multiple levels and the flexibility of virtualization, the proposed architecture is better suited to mobile networks than other schemes, as discussed in the paper.

- It can bring a Multi-RAT network comprising of LTE, 5G NR and, IEEE 802.11 under a unified SDN based control and management framework. It also enables a much simpler integration of IEEE 802.11 technologies with LTE/5G NR technologies through the VL, which virtualizes the underlying 5G NR, LTE and 802.11 PHY and MAC layer level resources in a unified manner for the higher level sub-plane functions like RLC and PDCP. This is not possible in the existing 3GPP LTE or 5G NR radio access network.
- The introduction of a VL between sub-planes (containing individual data plane layers) and associating the sub-planes through these virtualization layers does not require any deviation from the 3GPP standard. The resource abstractions are built over these layers by the VL as an additional mechanism without impacting their functionality. This differs from some of the schemes like Radiovisor, which may require changes in the protocol stack. Moreover, the changes do not impact UE protocol, and therefore, no changes are required in the UE to communicate with 5G/LTE/WLAN networks.
- VirtRAN offers a flexible mechanism for network virtualization and slicing, wherein virtualization (and consequently the slicing) can be performed at one or more different data plane layers (planes). Therefore, it is possible for a vendor/operator to choose the layers (planes) at which virtualization should be performed depending on the use cases that need to be supported. As shown through simulations, for the scenario depicted in Figure 10, virtualization and slicing can be done at the PDCP layer. An example of such a scenario is when an LTE network is to be used by two different operators to provide similar types of services. In such a scenario, virtualization can be done at the PDCP layer, and there may not be a need to virtualize it at a lower layer. Whereas for a different scenario, when the network slices are created to support different types of services, say uRLLC and eMBB (similar to the case in Figure 12), virtualization may be done at MAC layer. It is also possible to virtualize at multiple layers (planes) as shown in Figure 8.

IX. Conclusions

In this paper, we provide an overview of the existing solutions for RAN slicing in SDN/NFV based wireless networks. The paper also explores relevant ongoing standardization activities and describes their approach to slicing. We summarize and critique the available solutions while providing key insights into the requirements for slicing multi-RAN architectures. Finally, we propose and evaluate VirtRAN, an SDN/NFV based architecture for slicing multi-RAT RANs.

VirtRAN supports a unified framework for virtualization and slicing of multi-RAT wireless access networks. In order to enable an integrated virtualization framework for multi-RAT networks, it utilizes the concept of abstract network resources. These abstract network resources can be defined uniformly for different RATS and enable their integration under a single framework. It also provides a flexible approach towards network slicing. Network slicing under VirtRAN can be performed at one or more different data plane layers (planes), and depending on the requirement, one can choose the layers (planes) at which the network slicing is to be done. The architecture also provides a framework for recursive virtualization, which is essential for achieving policy-based resource allocation. Last but not least, even though the framework has been elucidated in the context of radio access networks, the concepts are generic and can be applied to other networks, such as mobile core networks. In the future, we would like to evaluate the applicability of this framework for the mobile core network. We would also like to evaluate it for the scenario when users are part of multiple slices.

X. Acknowledgement

This work has been supported by the Department of Telecommunications, Ministry of Communications, India as part of the indigenous 5G Test Bed project.

References

[1] K.-K. Yap, M. Kobayashi, R. Sherwood, T.-Y. Huang, M. Chan, N. Handigol, and N. McKeown, "OpenRoads: Empowering Research in Mobile Networks," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, pp. 125–126, 2010.

[2] G. Intelligence, "Global Mobile Trends: What is driving the mobile industry," 2018. [Online]. Available: https://www.gsmaintelligence.com/research/?file=8535289e1005eb248a54069d82ceb824&download.

[3] D. Kreutz, F. Ramos, P. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2014.

[4] E. Haleplidis, K. Pentikousis, S. Denazis, J. H. Salim, D. Meyer, and O. Koufopavlou, "RFC 7426: Software-Defined Networking (SDN): Layers and Architecture Terminology," 2015.

[5] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, 2008.

[6] ETSI GS NFV 003, "Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV," 2018.

[7] 3GPP TR 28.801 v15.1.0, "Study on Management and Orchestration of Network Slicing for Next Generation Network," 2018.

[8] E. Coronado, S. N. Khan, and R. Riggio, "5G-EmPOWER: A Software-Defined Networking Platform for 5G Radio Access Networks," *IEEE Transactions on Network and Service Management*, 2019.

[9] J. Schulz-Zander, L. Suresh, N. Sarrar, A. Feldmann, T. Hühn, and R. Merz, "Programmatic Orchestration of WiFi Networks," *Proceedings of USENIX Annual Technical Conference*, pp. 347–358, 2014.

[10] C. Hare, "Simple Network Management Protocol (SNMP)," 2011.

[11] R. Sherwood, M. Chan, A. Covington, G. Gibb, M. Flajslik, N. Handigol, T.-Y. Huang, P. Kazemian, M. Kobayashi, J. Naous, *et al.*, "Carving research slices out of your production networks with OpenFlow," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, pp. 129–130, 2010.

[12] E. Coronado, R. Riggio, J. Villa1ón, and A. Garrido, "Lasagna: Programming Abstractions for End-to-End Slicing in Software-Defined WLANs," in *2018 IEEE 19th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, pp. 14–15, IEEE, 2018.

[13] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software Defined Radio Access Network," in *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking*, pp. 25–30, ACM, 2013.

[14] S. Katti and L. E. Li, "Radiovisor: A Slicing Plane for Radio Access Networks," in *Presented as part of the Open Networking Summit 2014 ({ONS} 2014)*, 2014.

[15] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRAN: A Flexible and Programmable Platform for Software Defined Radio Access Networks," in *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*, pp. 427–441, ACM, 2016.

[16] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks—A Technology Overview," *IEEE Communications surveys & tutorials*, vol. 17, no. 1, pp. 405–426, 2014.

[17] 3GPP TS 23.214, " Architecture Enhancements for Control and User Plane Separation of EPC Nodes," 2017.

[18] 3GPP TS 23.501, "System Architecture for the 5G System," 2017.

[19] 3GPP TS 38.470 V15.6.0, "NG-RAN; F1 General Aspects and Principles," 2019.

[20] 3GPP TS 38.473 V15.7.0, "NG-RAN; F1 Application Protocol (F1AP)," 2019.

[21] 3GPP TS 38.463 V15.5.0, "NG-RAN; E1 Application Protocol (E1AP)," 2019.

[22] 3GPP TS 38.300 v15.6, " Overall description; Stage-2," 2019.

[23] O. N. Forum, "TR-502:SDN architecture issue 1.0," 2014. [Online]. Available: https://www.opennetworking.org/wp-content/uploads/2013/02/TR_SDN_ARCH_1.0_06062014.pdf.

[24] O. N. Forum, " TR-526:Applying SDN Architecture to 5G Slicing issue 1.0," 2016. [Online]. Available: https://www.opennetworking.org/wp-content/uploads/2014/10/Applying_SDN_Architecture_to_5G_Slicing_TR-526.pdf.

[25] A. Blenk, A. Basta, M. Reisslein, and W. Kellerer, “Survey on Network Virtualization Hypervisors for Software Defined Networking,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 655–685, 2015.
[26] O-RAN Alliance, “O-RAN: Towards an Open and Smart RAN,” 2018. [Online]. Available: https://www.o-ran.org/s/O-RAN-WP-FInal-181017.pdf.
[27] 3GPP TS 38.401 v15.5, “NG RAN - Architecture description,” 2019.
[28] “P1930.1 - Recommended Practice for Software Defined Networking (SDN) based Middleware for Control and Management of Wireless Networks.” https://standards.ieee.org/project/1930_1.html. Accessed: 2019-11-25.
[29] IEEE Computer Society LAN/MAN Standards Committee, “Part 11: Wireless LAN Medium-Access Control (MAC) and Physical Layer (PHY) Specifications,” tech. rep., 2016.
[30] IEEE Computer Society LAN/MAN Standards Committee, “IEEE Standard for Information technology-Telecommunication and information exchange between systems-Wireless Regional Area Networks (WRAN)-Specific requirements Part 22: Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Policies and Procedures for Operation in the TV Bands.,” tech. rep., 2011.
[31] D. Stanley, P. Calhoun, and M. Montemurro, “Control and Provisioning of Wireless Access Points (CAPWAP) Protocol Specification,” 2009.
[32] P. Calhoun, M. Montemurro, D. Stanley, *et al.*, “Control and Provisioning of Wireless Access Points (CAPWAP) Protocol Binding for IEEE 802.11,” *IETF RFC5416*, 2009.
[33] S. Kumar, D. Cifuentes, S. Gollakota, and D. Katabi, “Bringing Cross-layer MIMO to Today’s Wireless LANs,” in *ACM SIGCOMM Computer Communication Review*, vol. 43, pp. 387–398, ACM, 2013.
[34] K. Kozłowski, S. Kukliński, and L. Tomaszewski, “Open issues in network slicing,” *International Conference on the Network of the Future (NOF)*, 2018.
[35] “ns-3 Network Simulator.” [Online]. Available: https://www.nsnam.org.