An Architecture for Control Plane Slicing in Beyond 5G Networks

Rashmi Yadav, Rashmi Kamran, Pranav Jha, Abhay Karandikar

Abstract

To accommodate various use cases with differing characteristics, the Fifth Generation (5G) mobile communications system intends to utilize network slicing. Network slicing enables the creation of multiple logical networks over a shared physical network infrastructure. While the problems such as resource allocation for multiple slices in mobile networks have been explored in considerable detail in the existing literature, the suitability of the existing mobile network architecture to support network slicing has not been analysed adequately. We think the existing 5G System (5GS) architecture suffers from certain limitations, such as a lack of slice isolation in its control plane. This work focuses on the future evolution of the Third Generation Partnership Project (3GPP) 5GS architecture from a slicing perspective, especially that of its control plane, addressing some of the limitations of the 3GPP 5GS architecture. We propose a new network architecture which enables efficient slicing in beyond 5G networks. The proposed architecture results in enhanced modularity and scalability of the control plane in sliced mobile networks. In addition, it also brings slice isolation to the control plane, which is not feasible in the existing 5G system. We also present a performance evaluation that confirms the improved performance and scalability of the proposed system viz-a-viz the existing 5G system.

Keywords: Network slicing, beyond 5G network architecture, control plane slicing.

1. Introduction

The emergence of the Fifth Generation (5G) mobile networks enables a large variety of use cases and services. The Third Generation Partnership Project (3GPP) 5G System (5GS) categorizes its prominent use cases as Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communication (URLLC), Massive Machine Type Communication (mMTC), and Vehicle to Everything (V2X) (1), (2). Each of these use case categories is expected to support a different set of requirements. For example, eMBB use cases are expected to support very high data rates and low to high-speed mobility applications, while URLLC applications typically require very low latency and low to moderate data rates. Similarly, the broad characteristic of mMTC applications is to have low data rates along with very high connection density, while V2X applications require low latency support in high-speed mobility scenarios. Networks beyond the 5G era may need to support newer use cases such as holographic-type communications, tactile internet for remote operations, digital twin etc. Considering their diverse needs, each use case deserves a dedicated network infrastructure to efficiently serve the users. However, providing dedicated infrastructure to each of these use case categories may lead to an increase in Capital Expenditures and Operational Expenditures. Hence, the concept of network slicing is adopted in 3GPP 5GS to support the use-case category-specific requirements.

3GPP defines network slicing as “a paradigm where logical networks/partitions are created, with appropriate isolation, resources and optimized topology to serve a purpose or service category (e.g., use case) or customers (a logical system created on-demand)” (3). A network slice in 3GPP 5GS spans both horizontally as well as vertically, i.e., both Radio Access Network (RAN) and Core Network (CN) and also control and user (data) plane functions. In addition to the requirement to support multiple network slices with isolation between them, 3GPP 5GS needs to handle other slice-specific requirements too, as has been discussed in Section 16.3 of (4). We observe that one of these requirements, “Support for UE associating with multiple network slices simultaneously” (4), has a significant bearing on the control plane architecture of the 3GPP 5GS. The requirement mandates that “in case a UE is associated with multiple slices simultaneously, only one signalling connection is maintained”. It implies that the control plane functions especially that terminate UE signalling, e.g., a gNB-Centralized Unit-Control Plane (gNB CU-CP) function in RAN terminating Radio Resource Control (RRC) signalling, or an Access and Mobility Management Function (AMF) in CN terminating Non-Access Stratum (NAS) signalling, may have to support more than one slice concurrently. Therefore, achieving “isolation of slices” and having “slice-specific Network Functions (NFs)” in the control plane becomes particularly difficult. It should be noted that other 5G standards such as the one developed by O-RAN Alliance (5) also do not provide any guidance/resolution to this problem.

Network slicing in 5G networks has been an active field of research and here we provide a survey of the research work on this topic, especially those pertaining to isolation of slices. The work presented in (6) highlights the challenges of slice isola-
tion in the user (data) plane but there is no discussion on control plane slice isolation there. The authors in (7) investigate the challenges related to RAN slice design and implementation but do not discuss slice isolation. The authors in (9) guarantee functional and performance isolation of slices while allowing efficient use of resources in the RAN data plane but do not discuss isolation viz-a-viz control plane. A RAN slicing architecture with multiple sets of function splits and placements, which provide isolation among slices, is proposed in (9). However, it neither discusses control plane slice isolation nor does it discuss core network slicing. Another work (10) proposes a flexible RAN architecture with a Medium Access Control (MAC) scheduler to abstract and share physical resources among slices. In (11), Software Defined Network (SDN) enabled resource allocation framework is proposed and in (12), a network slicing framework for end-to-end Quality of Service (QoS) with a dynamic radio resource slicing scheme is proposed. In (13), an architecture for cloud-network slicing concept and realization of the slice-as-a-service paradigm is presented. It is designed to consider modularity and multi-domain dynamic operation as key attributes.

As can be discerned from the literature survey presented above, most existing works focus only on the slicing aspects of the user (data) plane and to the best of our knowledge, no prior work on control plane slice isolation is found. Moreover, the architectural mechanism to enable slice isolation in the control plane of 3GPP 5GS has neither been discussed in the standards nor in the research work.

Therefore, in order to achieve slice isolation in the mobile network control plane, we propose a new mobile network architecture in this paper. The proposed architecture improves the slice-specific design of the mobile network control plane and facilitates slice isolation therein. It is an extension of our earlier 5G-Serv architecture [16]. The 5G-Serv did not explore slicing aspects, which has been addressed in this extension. In addition, we have also done a detailed performance evaluation of the control plane in the proposed architecture in a sliced environment and compared it with the 3GPP 5GS architecture. The performance evaluation focuses on slice-wise session establishment rate, resource utilization, scalability and modularity of the control plane. We demonstrate that the proposed architecture achieves improved control plane performance in a sliced environment as compared to the 3GPP 5GS architecture. An additional benefit of the proposed architecture is its simplified end-user signalling viz-a-viz the 3GPP 5GS.

The rest of the paper is organized as follows: Section 2 discusses the architectural details of the proposed slice-specific control plane architecture. The procedures involved in the proposed extension are detailed in Section 3. Section 4 provides the system model. The performance evaluation is covered in Section 5 while we conclude in Section 6.

2. Proposed Architecture for Control Plane Slicing

In the 3GPP 5GS, one of the issues behind the problem of slice isolation in the control plane is the termination (placement) of UE signalling functionality within the control plane function, e.g., in gNB-CU-CP or AMF. Hence, the separation of UE signalling handling functionality from the control plane may help us in solving the problem of slice isolation in the control plane. The idea to have separate NFs for UE signalling handling, separate from the existing control plane functions, was proposed in our earlier work on 5G-Serv (details available in [14]). However, it was not analysed in the context of a sliced network. The architecture of the proposed mobile network is divided into slice-specific control plane and data plane NFs in RAN and CN as shown in Fig. 1.

In our proposal, each plane has slice-specific NFs in RAN and core network. Besides, a few functions are slice-specific or common to a set of slices such as UE signalling service function (comprising RRC and NAS signalling). In this work, we consider two network slices for simplicity. However, the proposed architecture can easily employ more than two slices. Various components of the proposed architecture are explained in the following sections:

2.1. Control Plane

The control plane functions in RAN and CN of the proposed architecture are named as the RAN controller and the CN controller, respectively. The RAN controller is responsible for resource allocation and data (user) plane control functionality in RAN, whereas CN controller is responsible for the same functionality in CN.

In the 3GPP 5GS, the gNB-CU-CP, the de facto RAN control plane function, broadly contains the following functionalities: gNB-DU control, gNB-CU-UP control, RRC protocol, Radio Resource Management (RRM) and Next Generation Application Protocol (NGAP) functionalities. We propose to change the placement of some of these functionalities to simplify the RAN control plane function for end-to-end slicing. The above-mentioned functionality can broadly be divided into two classes: (i) UE-specific control/signalling functionality, e.g., UE-specific RRC protocol functionality and (ii)
RAN user plane control functionality, e.g., gNB-DU/gNB-CU-UP control functionality. The UE-specific signalling functionality in gNB-CU-CP, responsible for terminating RRC protocol signalling with UEs, is moved out of the gNB-CU-CP and is transposed to a new UE signalling service function in the proposed architecture. Also, UE-specific NGAP message handling, to carry NAS messages between Access and Mobility Management Function (AMF) and gNB-CU-CP in the 3GPP 5GS, is also removed from the gNB-CU-CP. After the relocation of the UE-specific signalling handling functionality from gNB-CU-CP to UE Signaling Service Function, only the RAN user plane control functionality remains part of gNB-CU-CP and simplifies the overall gNB-CU-CP design. This simplified gNB-CU-CP is rechristened as RAN controller in the proposed architecture.

Similarly, UE-specific signalling functionality, e.g., NAS signalling handling, UE authentication etc. are moved out of the CN control plane functions like AMF, Session Management Function (SMF), Authentication Server Function (AUSF) and are placed in the UE signalling service function. The remaining user plane control functionality in CN control plane is rechristened as CN controller, which is considerably simpler compared to the conventional CN control plane in 3GPP 5GS. The modified RAN and CN controllers can communicate through an inter-controller interface similar to the existing NGAP interface.

Note that the modified RAN and CN controllers (in the proposed architecture) do not contain the UE signalling functionality and no longer terminate the UE signalling. This architectural change allows the controllers to be slice-specific, i.e., every network slice can have its own RAN and CN controllers removing the constraint for the control plane to necessarily support more than one slice in 3GPP 5GS.

2.2. Data (User) Plane

The user plane is responsible for the transfer of data through the mobile network. There are no changes in the user plane functionality in the proposed architecture over the 3GPP 5GS; these remain the same as the user plane functions in the 3GPP 5GS. The gNB-CU-UP comprises Service Data Adaptation Protocol (SDAP), General Packet Radio Service (GPRS) Tunneling Protocol (GTP), and Packet Data Convergence Protocol (PDCP) layers, and the gNB-DU has Radio Link Control (RLC), MAC, and Physical (PHY) layers. gNB-DU and gNB-CU-UP are together termed as RAN-Data Plane (RAN-DP) or gNB-Data Plane (gNB-DP). RAN-DP (gNB-DP) in RAN and the Core Network Data Plane (CN-DP (User Plane Function (UPF))) in CN maybe slice specific, i.e., each logical network (slice) has its RAN-DP and CN-DP (UPF) functions.

2.3. UE Signalling Service Function

UE signalling service function exchanges signalling messages such as RRC/NAS messages with UEs. This is a new function defined as part of the proposed architecture. The CN control plane functionality in 3GPP 5GS such as NAS signalling termination as part of AMF or UE authentication functionality as part of AUSF are moved from the CN control plane functions to the UE signalling service function in the proposed architecture as is the RRC signalling handling functionality from gNB-CU-CP. These UE signalling service functions can either be slice-specific or common to a set of slices. It is possible to have more than one UE signalling service function in the network for reasons such as load balancing and distribution of functionality across them.

2.4. Interfaces

Fig 2 shows proposed/modified interfaces in the proposed architecture. gNB-CU-CP has been segregated into two different entities as RAN controller and UE signalling service function. Conventionally, F1-C is the interface between gNB-DU and gNB-CU-CP and it carries UE-specific messages. However, a modified F1-C (F1-C’) interface between RAN-DP and RAN Controller does not carry UE-specific control messages and information elements. Besides, a new interface, F1″ is proposed between RAN-DP and UE signalling service function, which now carries UE-specific RRC/NAS messages. An important consequence of the creation of UE signalling service functions, separate from the control plane functions is that UE-specific signalling messages are another form of data passing through the user plane. Hence the proposed F1” interface is similar to the F1-U interface of the 3GPP 5GS. Now, UE can exchange signalling message with UE signalling service function via RAN-DP over F1″ interface. Controller-service function interface exists between the RAN controller and the UE signalling service function. Inter-controller interface is a new interface which can be based on the existing NGAP interface.

The proposed architecture is validated and elucidated further through an example of Protocol Data Unit (PDU) session establishment call flow in the next section.

3. PDU Session Establishment Call Flow

3.1. 3GPP 5GS

In this section, we detail the call flow involved in establishing a PDU session for a UE to access two network slices in the case of 3GPP 5GS, as shown in Fig. 3 (referred from Section 4.3.2.2 of [15]).

- The UE initiates the process by sending a PDU session establishment request to the AMF (message 1).
- The AMF selects the appropriate SMF (SMF-1 in this case), sends a create PDU session context request to SMF-1 (message 2) and receives a response from SMF-1 (message 3).
• SMF-1 configures UPF-1 through an N4 session establishment request and response (messages 4 and 5).

• SMF-1 communicates with AMF using N1N2 message transfer (messages 6 and 7).

• AMF forwards the N2 PDU session request to the RAN-Centralized Unit (CU)-1 (message 8) and receives its response in message 12.

• RAN-CU-1 sends a DRB configuration message to RAN-Distributed Unit (DU)-1 (message 9), and a PDU session establishment accept message to the UE (message 10).

• The RRC reconfiguration message occurs between UE and RAN-CU-1 (message 11).

• AMF sends PDU session update context request to SMF-1 (message 13) and receives its response in message 16.

• SMF-1 modifies the N4 session in messages 14 and 15.

• The PDU session is now established, allowing the UE to access slice 1. A similar process is followed for accessing the second slice as shown in messages 17 to 32.

3.2. Proposed architecture

This section discusses the call flow of PDU session establishment for a UE to access two network slices in the proposed architecture (as shown in Fig. 4 to understand its working. Besides, we also compare it with the PDU session establishment call flow in the 3GPP 5GS under the same scenario of two slices, as shown in Fig. 3 (referred from Section 4.3.2.2 of (15)).

At the start, UE sends a PDU session establishment request to access a network slice (say slice 1). Based on the received request, the UE signalling service function (common to both the slices in our case) forwards the request to the respective CN controller. Further, the CN controller selects the corresponding CN-DP (UPF) for session establishment, and N4 session is established at the CN-DP (UPF). In Fig. 4, the call flow is shown for a UE connecting two network slices in which UE first connects to slice 1 and then to slice 2. Details of message sequences in the call flow are as follows:

• UE sends a PDU session establishment request as a NAS message to the UE signalling service function (message 1).

• Based on the received request for a particular slice, the UE signalling service function selects the slice-specific controller (say CN controller-1 for the first slice) in the core network and sends a message to create a PDU session context (message 2).

• Accordingly, the CN controller-1, which is specific to the slice 1, configures CN-DP (UPF)-1 (message 3).
The CN controller-1 informs the RAN controller-1 (which is also slice-specific) about the PDU session setup (message 4) on the inter-controller interface.

Subsequently, the RAN controller-1 establishes a PDU session on the RAN-DP-1 by sending a Data Radio Bearer (DRB) configuration message to RAN-DP-1 (message 5). It also notifies about DRB and PDU session establishment (message 6) to the UE signalling service function (after the PDU session is established).

UE signalling service function sends the RRC reconfiguration message to the UE (message 7).

RAN controller-1 sends PDU session context update to CN controller-1 (message 8).

CN controller-1 sends N4 session modification command to CN-DP (UPF)-1 (message 9).

PDU session is now established for UE to access slice 1. The same message sequence is followed for UE to access the second slice, as shown in the second part of Fig. 4.

On comparing the message sequences for accessing the network slices in the case of the 3GPP 5GS (Fig. 3) and the proposed system (Fig. 4), we find that there is a reduction in the total number of messages by employing the proposed architecture as compared to the 3GPP 5GS. For instance, in the 3GPP 5GS procedure, there are request/response messages sent and received between AMF and SMF for creating and updating a PDU session context (messages 2,3,13,16,18,19,29 and 32). Likewise, request/response messages were sent and received between SMF and UPF (messages 4,5,14,15,20,21,30 and 31) for N4 session establishment and modification. In contrast, responses from the CN controller to the UE signalling service function (for the PDU session create context) and the RAN controller (for the PDU session context update) are not required. In a similar way, responses from CN-DP (UPF)-1 to the CN controller (for N4 session establishment and modification) are not required. N1N2 messages (messages 6,7,22 and 23), which are communicated between AMF and SMF in the 3GPP 5GS, are removed completely from the proposed call flow as the RAN controller and the CN controller can communicate directly through an inter-controller interface. Overall, the message sequence for PDU session establishment (for two slices) for the proposed scheme is simplified as compared to the 3GPP 5GS, which enables enhanced modularity in the procedures of the proposed system.

4. System Model

In this section, we describe the system model of the proposed architecture by considering the example of PDU session establishment call flow. We have used Performance Evaluation Process Algebra (PEPA) (16), a formal high-level language for modelling distributed systems and their evaluations.

The modelling of the proposed call flow (Fig. 4) is provided in Table I. Various NFs, e.g. UE, RAN-DP-1 (DP-1), UE signalling service function (USSF), RAN controller-1 (RANC-1), CN controller-1 (CNC-1), and CN-DP (UPF)-1 for the proposed architecture are modelled as PEPA modules. The NF’s states are defined with the corresponding NF name and a number (NF1) (e.g., Ranc1 indicates the first state for the RAN controller). Further, the action types are denoted in lowercase and subscripts are added to specify the detail of the defined actions (actiontype). For example, request and reply for any service, e.g. PDU session create context, can be specified as reqsc and repsc, respectively. Each action type is associated with a specific rate value, \( r \). The rates (number of actions performed per unit time) model the expected duration of a specific type of action in the PEPA component and are taken as reference from (17), (18) and (19).

Let us consider an NF, for example, CN-DP (UPF)-1, to understand the modelling of the system NF. Various messages (actions) are associated with this NF (CN-DP (UPF)-1) during the
Table 1: System modelling for the proposed architecture

<table>
<thead>
<tr>
<th>PEPA Modelling of NFs</th>
<th>Code Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UE</td>
<td>$Ue_1 = \text{get}_r, Ue_2 = \text{req}_r, Ue_3 = \text{req}_r$</td>
</tr>
<tr>
<td>NF</td>
<td>$Ue_4 = \text{req}_r, Ue_5 = \text{req}_r, Ue_6 = \text{req}_r$</td>
</tr>
<tr>
<td>DP-1</td>
<td>$Dp_1 = \text{db}, Dp_2 = \text{setup}_r, Dp_3 = \text{update}_r$</td>
</tr>
<tr>
<td>RANC-1</td>
<td>$Ranc_1 = \text{setup}_r, Ranc_2 = \text{update}_r$</td>
</tr>
<tr>
<td>USSF</td>
<td>$Ussf_1 = \text{req}_r, Ussf_2 = \text{req}_r, Ussf_3 = \text{req}_r$</td>
</tr>
<tr>
<td>CNC-1</td>
<td>$Cnc_1 = \text{setup}_r, Cnc_2 = \text{req}_r, Cnc_3 = \text{req}_r$</td>
</tr>
<tr>
<td>CN-DP (UPF)-1 NF</td>
<td>$Upf_1 = \text{setup}_r, Upf_2 = \text{req}_r, Upf_3 = \text{req}_r$</td>
</tr>
<tr>
<td>UE Processor</td>
<td>$Uep_1 = \text{setup}_r, Uep_2 = \text{req}_r$</td>
</tr>
<tr>
<td>DP-1 Processor</td>
<td>$Dpp_1 = \text{setup}_r, Dpp_2 = \text{req}_r$</td>
</tr>
<tr>
<td>RANC-1 Processor</td>
<td>$Rancp_1 = \text{setup}_r, Rancp_2 = \text{req}_r, Rancp_3 = \text{req}_r$</td>
</tr>
<tr>
<td>USSF Processor</td>
<td>$Ussf_1 = \text{setup}_r, Ussf_2 = \text{req}_r$</td>
</tr>
<tr>
<td>CNC-1 Processor</td>
<td>$Cncp_1 = \text{setup}_r, Cncp_2 = \text{req}_r$</td>
</tr>
<tr>
<td>CN-DP (UPF)-1 Processor</td>
<td>$Upfp_1 = \text{setup}_r, Upfp_2 = \text{req}_r$</td>
</tr>
<tr>
<td>System Equation</td>
<td>$(((Uep_1 \oplus Dpp_1 \oplus \text{get}_r \oplus \text{setup}_r) \oplus \text{get}_r \oplus \text{setup}_r) \oplus \text{get}_r \oplus \text{setup}_r) \oplus \text{req}_r \oplus \text{req}_r$</td>
</tr>
<tr>
<td>Variables</td>
<td>$S_1 = \text{req}_r, S_2 = \text{null}, S_3 = \text{null}, S_4 = \text{null}, S_5 = \text{null}, \phi = \text{null}$</td>
</tr>
</tbody>
</table>

This section presents the performance evaluation of the proposed end-to-end slicing network solution. We have created two slices in each of the 3GPP 5G S's and in the proposed architecture and then we have evaluated the performance of the session establishment procedure in a sliced environment. We
A few lines of text are missing from the image, making the content incomplete. However, the visible text seems to be related to network function (NF) modeling and performance evaluation in 3GPP 5GS architecture. The text includes equations and processes related to the configuration and performance of NFs, with expressions involving processor utilization, configuration, and performance measures.

The text mentions the evaluation of established sessions, the comparison of 3GPP 5GS with the proposed architecture, and the calculation of processor utilization and performance measures. It also briefly describes the use of configuration parameters and their impact on system performance.

The text appears to be discussing the mathematical formulation of performance metrics in 3GPP 5GS, with a focus on how to evaluate the performance of NFs and the overall system in a sliced environment. It mentions the use of equations to calculate processor utilization and performance metrics. The system is evaluated through various performance measures such as the number of sessions established per unit time, Average Response Time (ART), and processor utilization in a sliced environment.

The text also highlights the importance of comparing the 3GPP 5GS with the proposed architecture, noting that the scalability of the network is a key factor in this comparison. It mentions the use of equations to calculate the productivity of a system at scale and the average number of PDU sessions established at scale, with a focus on how these metrics are determined through response time performance evaluations.

The text concludes with a discussion on how the configuration parameters affect the system's performance and how to optimize the network for better scalability and efficiency.
users and for the proposed architecture is at 20,000 users for basic configuration, shown in Fig. 5. Similarly, Fig. 6 shows that for scaled configuration, the 3GPP 5GS saturates at 20,000 users, while the proposed architecture saturates at 46,000 users. Here saturation point is the maximum number of users that can be served by the system. This saturation point corresponds to the saturation of processor utilization (discussed next).

The processor utilizations (for all the NFs of slice-1) of the 3GPP 5GS architecture with the scaled configuration of $(N_{dp1}, N_{dp2}, N_{amf1}, N_{smf1}, N_{smf2}, N_{upf1}, N_{upf2}) = (1,1,1,1,1,1,1)$ and for the 3GPP 5GS architecture with the basic configuration of $(N_{du1}, N_{du2}, N_{smf1}, N_{smf2}, N_{upf1}, N_{upf2}) = (1,1,1,1,1,1,1)$. 

Based on the results obtained for the slice-wise session establishment rate, ART and processor utilization (from the PEPA-based simulation and modelling), scalability is evaluated (from Equation 1) and plotted in Fig. 9. We consider the same two configurations $m_1$ and $m_2$ as referred to above for the estimation of scalability. The same saturation points can be observed in Fig. 9. The 3GPP 5GS saturates at 8,000 users, while the proposed architecture saturates at 20,000 users for basic configuration of $(N_{dp1}, N_{dp2}, N_{amf1}, N_{amf2}, N_{rn1}, N_{rn2}, N_{lapf1}, N_{lapf2}) = (1,1,1,1,1,1,1,1)$ and for the 3GPP 5GS architecture with the basic configuration of $(N_{du1}, N_{du2}, N_{amf1}, N_{amf2}, N_{smf1}, N_{smf2}, N_{upf1}, N_{upf2}) = (1,1,1,1,1,1,1,1)$.
Acknowledgment

We acknowledge the Ministry of Electronics and Information Technology (MeitY), Govt. of India for supporting the project.

References