Applying SDN to Mobile Networks: A New Perspective on 6G Architecture

ANONYMOUS AUTHOR(S)

The upcoming Sixth Generation (6G) mobile communications system envisions supporting a variety of usage scenarios with differing characteristics, e.g., immersive communication, hyper reliable and low-latency communication, ultra massive connectivity, ubiquitous connectivity, haptic communications etc. To accommodate such diverse scenarios, the 6G system (6GS) architecture needs to be scalable, modular, and flexible. In this article, we identify some limitations of the Third Generation Partnership Project (3GPP) defined Fifth Generation System (5GS) architecture, especially that of its control plane. Further, we propose a novel architecture for the 6GS employing Software Defined Networking (SDN) technology to address these limitations. Among the different functionalities of the 5GS control plane, two key functionalities are the signalling exchange with end user devices (e.g., for user registration and user authentication) and control of user plane functions. We propose to move the "signalling handling functionality" out of the mobile network control plane and treat it as user service, i.e., as payload or data. This proposal results in an evolved service-driven architecture for mobile networks where almost all communication with an end user (and device), including the signalling exchange, is treated as service. We show that the proposed architecture brings increased simplicity, modularity, scalability, flexibility and security to its control plane. To demonstrate the advantages of the proposed architecture, we also compare its performance with the 5GS using a process algebra-based simulation tool.

ACM Reference Format:
Anonymous Author(s). 2024. Applying SDN to Mobile Networks: A New Perspective on 6G Architecture. 1, 1 (March 2024), 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The notable rise in the diversity of use cases has paved the way for the continued evolution of mobile networks. The upcoming 6th Generation (6G) Mobile Communication System is envisioned to support new use cases such as holographic-type communications, tactile internet, intelligent operation networks, digital twin, and Industrial Internet of Things (IIoTs) with cloudification [1]. It is also foreseen that there will be a large number of connected users in the 6G era enabled by usage scenarios like 'Ubiquitous Connectivity' and 'Massive Communication' [2]. A scalable, flexible and modular network architecture is one of the essential ingredients towards tackling the diverse usage scenarios and the anticipated massive connectivity in 6G networks. These architectural characteristics would be particularly important for the network control plane which would bear the brunt of the enormous signalling load generated by the huge number of users [3].

Third Generation Partnership Project (3GPP) adopted technologies such as Network Function Virtualization, Control and User Plane Separation, and Network Slicing for Fifth Generation System (5GS), which resulted in improved scalability and flexibility of 5GS over the previous generation mobile communications systems such as Fourth Generation System (4GS). However, there is scope for further improvement in mobile network architecture, especially that of its control plane through the application of Software Defined Networking (SDN) technology. A survey of the existing research related to SDN-based enhancements in the mobile network

ACM XXXX-XXXX/2024/3-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

control plane is presented next. The work in [4] proposes a centralised control plane for multi-Radio Access Technology (multi-RAT) Radio Access Network (RAN) to enhance the simplicity and flexibility of the network. Relocation of the control plane functionality of RAN to the Core Network (CN) to reduce the signalling cost between RAN and core has been discussed in [5]. Authors in [6] proposed a decentralized control plane architecture for the 5GS with independent control functions for different control events for flexible and scalable networks. An SDN architecture where a middle cell and a middle cell controller are introduced between the macro cell and the small cell to reduce the control overhead of the macro cell and to address the scalability problems is proposed in [7]. In [8], authors proposed a new 5GS core architecture based on the SDN concept. They introduced a centralised SDN controller for easier and more flexible management of the user plane. In [9], a hierarchical control plane is designed to lighten the load of the controller. It focuses on the vertical scalability of the control plane. In [10], a scalability metric for the SDN control plane is proposed. Besides, a comparison between different SDN architectures is analysed via mathematical methods. In [3], authors propose to process a subset of signalling messages within the user plane (data plane). In addition, there is a vast amount of literature on SDN-based network architectures, albeit unrelated to mobile networks [11], [12].

To summarize, current research in the context of the application of SDN technology to mobile networks mainly focuses on the centralized or distributed architecture of the control plane for reduced control overheads or scalability purposes. However, to the best of our knowledge, there is a limited discussion/rethink on certain other aspects of network architecture, such as, what functionality should constitute the mobile network control plane within an SDN-based architecture, is the network control plane right place for 'end user signalling handling' functionality in such an architecture? Should 'Non-Access Stratum (NAS) messages' be handled by CN control plane functions such as Access and Mobility Management Function (AMF) or should this functionality be moved out of AMF? Should the user authentication function (Authentication Server Function (AUSF) in 5GS) be part of the CN control plane? These questions assume even more importance in the upcoming 6G era, when the increased end-user signalling load due to a surge in the number of users has the potential to over-burden the network control plane.

In order to bring in additional enhancements to mobile network architecture, especially to its control plane, we propose to altogether separate end user device (User Equipment (UE)) signalling handling from the control plane functions. In a significant departure from the existing cellular networks, the proposed architecture views 'UE signalling' as payload, i.e., a form of data traversing through the cellular network, not much different from other types of data such as 'video streaming' or 'web browsing'. We analyse and evaluate the proposed architecture using Performance Evaluation Process Algebra (PEPA) [13], a formal language used to model distributed systems. We also provide a comparative analysis of the proposed architecture and the 5GS architecture through example call flows for Protocol Data Unit (PDU) session establishment and UE mobility procedures. We demonstrate a significant reduction in the number of control messages exchanged in the proposed architecture along with an improvement in network scalability.

The rest of the paper is organised as follows: Section 2 provides limitations of the 3GPP 5GS architecture. Section 3 provides an overview of the proposed architecture and highlights its advantages. Section 4 includes an information flow comparison of the 5GS and proposed architecture for PDU session establishment and mobility procedures. Section 5 describes the system model using PEPA. Section 6 covers the performance analysis. Section 7 provides the conclusion and information on the future work.

2 LIMITATIONS OF 3GPP 5GS ARCHITECTURE

In this section, we have captured some of the limitations of the 3GPP 5GS architecture especially that of its control plane. Although there can be other limitations too say pertaining to radio technology, etc., those are not discussed here.

2.1 Tight coupling of user plane control and UE signalling in control plane

The 5GS architecture supports the control and user plane separation. Among other functionalities, the 5GS control plane performs user plane control (network resource control, e.g., setting up data path through the user plane) and UE signalling handling functionalities (e.g., NAS/RRC (Radio Resource Control) message exchange with UEs). There is a tight coupling between these two categories of functionalities, i.e., between user plane control and UE signalling handling and certain CN (e.g., AMF) and RAN (gNodeB-Centralized Unit-Control Plane (gNB-CU-CP)) control plane functions in the 5GS perform both. A detailed description of control plane functionality is provided in [14]. This may lead to issues of control plane scalability due to the high signalling load caused by the presence of a very large number of users in future networks. As demonstrated here, decoupling of UE signalling handling functionality from user plane control functionality may lead to a more modular and scalable network architecture.

2.2 Limited alignment with SDN paradigm

SDN is a networking paradigm which separates the control plane of a network from its user (data) plane and centralizes the network's intelligence in the control plane. Although there are differing views in industry/academia on how to define an SDN-based network architecture, we can still discern a broad agreement on the topic [6], [15], [16]. The 5GS architecture incorporates the concept of SDN, resulting in architectural features such as the separation of the user plane from the control plane [14]. However, closer observation shows that the 5GS architecture does not align completely with the SDN paradigm. Besides controlling the user plane, the 5GS control plane also exchanges signalling messages with UEs to provide services such as authentication and also collect service requirements, e.g., requirements for PDU connectivity service. The functionality of signalling exchange with UEs may fit better within the service plane instead of the control plane in an SDN based mobile network [17].

2.3 Non-uniform handling of services

Services in the 5GS can be categorized into the following two types:

(1) Application-based services such as media streaming services, Internet Protocol (IP) multimedia subsystem services, mission-critical services, Multicast/Broadcast Services (MBS), etc.
(2) Other than these application-based services, the 5GS also provides services such as initial access, registration, authentication, PDU connectivity (connectivity to data networks), and connected mode mobility support. Such services can be called built-in (or internal) network services.

The two categories of services (application based services and built-in network services) are enabled differently in the 5GS. As Application (Service) Functions (AFs) are independent and decoupled from the CN and RAN functions of mobile networks, they access the control plane functions of the mobile CN over a standardized interface to enable service delivery through the user plane. However, the delivery of built-in services is tightly integrated within the control plane of the 5GS (both RAN and CN).

2.4 Inconsistent support for principle of "separation of concern"

Even though 5GS has separate control and user plane functions, an altogether clean separation of functionalities between the two planes is missing. For example, a glaring anomaly is the transfer of the Short Message Service (SMS), a form of user data, to the UEs via control plane functions like AMF and gNB-CU-CP. SMSs are delivered using NAS signalling messages unlike other user data typically delivered via PDU sessions. A similar but contrasting example is that of Access Traffic steering, Switching, and Splitting (ATSSS) functionality at User Plane Function (UPF). To aid the ATSSS functionality, ‘Measurement Assistance Information’, a type of signalling information is exchanged between the UE and the Performance Measurement Function (PMF), a sub-function within UPF. Even though ‘Measurement Assistance Information’ is a type of signalling information, it is exchanged via a PDU session (i.e. via the user plane functions solely) between the UE and the PMF. The mechanism is different from how other signalling information such as “radio measurement reports” to support the "mobility procedure" is exchanged, i.e., via dedicated signalling paths. To summarize, the 5GS does not use regular paths for both data as well as signalling exchange in certain scenarios bringing inconsistency to the architecture.

2.5 Complex protocols between control plane and user plane

The 5GS control plane architecture impacts the interface design (protocols) between the control and user planes. For instance, F1 Application Protocol (F1AP) is the protocol used on the interface between the RAN control plane (gNB-CU-CP) and the RAN user plane (gNB-Distributed Unit (gNB-DU) or RAN-DU). It is primarily used to configure gNB-DU but also carries RRC (UE signalling) messages for UEs as the control plane also handles UE signalling. Integrating both these types of functionalities in a single protocol results in a relatively complex communication protocol between gNB-CU-CP and gNB-DU.

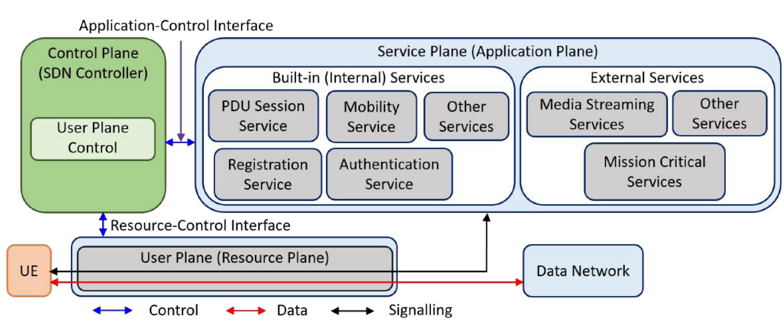

Fig. 1. Proposed SDN-based architecture for 6G system [18].

3 PROPOSED ARCHITECTURE FOR 6G SYSTEM (6GS)

This section presents the proposed architecture, which addresses the architectural limitations of the 5GS (as discussed in Section 2) and highlights a few other advantages. In the proposed work, we aim to separate the UE signalling handling from the control plane and treat them (UE signalling) as a service (data) to the user to enhance the scalability, modularity and flexibility in the mobile network control plane. The proposal results in an evolved service-driven architecture for mobile networks where almost all communication with a user (and its device), including the signalling exchange, is treated as service. With the proposed separation, the control plane becomes quite thin and is left with only the user plane control functionality, as shown in Figure 1. The UE signalling handling functionality is moved out of the control plane to the service plane. The service plane consists of various in-built and external service functions, as shown in

Figure 1, such as the PDU session service function (handles PDU session establishment and management providing PDU connectivity service), mobility service function (responsible for handling UE mobility), registration service function (handles UE registration with the network), authentication service function (manages UE authentication) and a few others. Due to the reorganisation of the architecture, it offers many architectural and performance advantages, discussed next. Please note that there may be separate controllers in the CN and RAN, as shown in Figure 3. Further, the proposed architecture's user or resource plane may remain the same as the 3GPP 5GS with only minor changes.

3.1 Advantages of the proposed 6GS architecture

This section highlights a few advantages of the proposed work. Segregation of UE signalling handling functionality from the control plane **simplifies the control plane** making it thinner viz-a-viz the 5GS control plane and enhancing its scalability and modularity.

The reorganised architecture also **aligns well with the SDN paradigm** as the control plane is redesigned to perform only user plane control functionality as discussed in Section 2.2. The proposed architecture also allows internal (or built-in 5GS) services to be treated the same way as external application-based services, leading to **uniform handling of all types of services**.

Further, this proposal results in the simplification of the control messages. For instance, the number of sessions management-related messages is reduced due to the setup of a direct path between UE and the service function (such as PDU session service function and mobility service function (detailed in Section 4.2)), leading to **simplified information(call) flows**. Also, the number of hops between the RAN controller and the CN controller in the proposed architecture is less than the corresponding entities in 5GS, i.e., between gNB-CU-CP and the Session Management Function (SMF), respectively, which further results in the performance improvement in terms of control plane latency and resource utilisation. Transposition of UE signalling handling functionality to functions in service plane **simplifies the protocols** between the control pane and the user plane such as Next Generation Application Protocol (NGAP) between the CN control plane (AMF) and RAN (gNB) and F1AP between the RAN control plane (gNB-CU-CP) and the RAN user plane (gNB-DU).

The proposed architecture also utilizes the **principle of separation of concern** and there is a clear-cut demarcation between the user and the control plane functionality here unlike the earlier generation systems including the 5GS. For example, the control plane of the proposed architecture does not directly participate in transfer of user data such as SMS, as is the case with 5GS. These are handled by the service functions and the user plane functions.

The 5GS uses the same type of signalling messages for all use cases. However, it is possible to have different signalling requirements for different use cases, e.g., the Internet of Things (IoT) and human user devices. The proposed architecture may support this requirement by employing **use case specific signalling** service functions. Our proposal can also support **flexible function deployment and chaining** as various service functions, such as the PDU session service function, mobility service function, registration service function, and authentication service function, can be placed flexibly and chained together to serve UEs.

An additional advantage of the proposed architecture towards network access security is presented here. 3GPP specification [19] highlights the exposed AMF which is vulnerable to replay attacks of NAS signalling messages between the UE and AMF (control plane of the CN). In a similar way, [20] presents the exposed RAN which is susceptible to replay attacks via RRC signalling messages exchanged between the UE and gNB-CU-CP, the control plane of 5G RAN, as the Uu interface also carries sensitive RRC signalling. Further, the European Union Agency for Cybersecurity (ENISA) [21], in its report, highlights that the N2 interface between the 5GS RAN and AMF is a target for attackers since they carry sensitive signalling between the RAN and the CN. These scenarios highlight the "access security threats" posed by UE signalling to the control plane of the 5G network, AMF and gNB-CU-CP. Since UE signalling handling is segregated from the control plane (of RAN and CN) in the proposed architecture and is terminated to a separate signalling server, it leads to the possibility of localizing the attack originating from a UE within the signalling servers without compromising the network control plane, where the logical control and management of RAN and CN are located. This segregation may allow us to improve the network access security in future mobile networks.

Please note that there is no impact on the UE both with respect to signalling exchange as well as data transfer in the proposed architecture viz-a-viz the 5GS. The signalling protocol between the UE and the network remains the same as the 5GS. The impact is only on the network architecture and the message flow between different functions on the network side.

4 SIGNALLING AND CONTROL INFORMATION FLOW COMPARISON

In this section, we compare the signalling and control information flow of the proposed architecture and the 5GS architecture. We consider the PDU session establishment and (user) mobility service examples to differentiate the working of the 5GS and the proposed architectures.

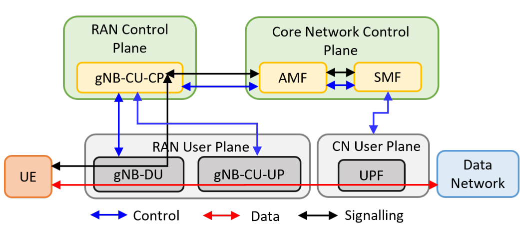

Fig. 2. Network entities, UE signalling and control message flow for PDU session establishment in 5GS.

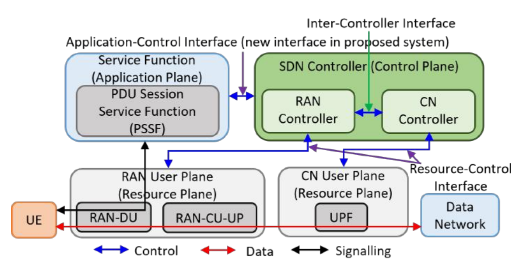

Fig. 3. Network entities, UE signalling and control message flow for PDU session establishment in the proposed architecture.

4.1 PDU session establishment

Figure 2 and Figure 3 show the entities involved in control flow and signalling exchange for PDU session establishment for the 5GS and the proposed architecture, respectively. In 5GS, messages are exchanged between UE and SMF for PDU session-related signalling via RAN (it requires gNB-DU and gNB-CU-CP) and AMF. However, signalling messages are directly exchanged between UE and the service function (PDU session service function (PSSF)) via RAN (it requires only RAN-DU) in the proposed architecture, as shown in Figure 3. It implies that in the 5GS, signalling messages pass through multiple hops. In contrast, the number of hops is reduced in the proposed architecture. Further, the control plane collects all requirements from UE via the application-control interface and establishes the PDU session.

The complete message sequences for establishing PDU sessions for the 5GS are detailed in [18] while simplified call flow for the proposed architecture is shown in Figure 4[1]. Please note that the controllers do not require response messages from the resource (user) plane,

[1] In call flows and simulations, only those messages are considered and compared which are different in proposed and 5GS architectures

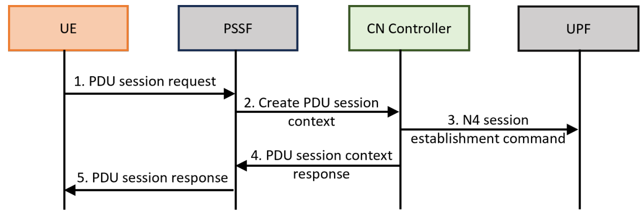

Fig. 4. PDU session establishment procedure in the proposed architecture.

as the controller knows about user plane resource information; it handles resource decision-making. Therefore, the proposed architecture eliminates many such messages. For example, the N4 session modification request and response are exchanged between SMF and UPF in 5GS architecture [18], while the session establishment command (message 3 in Figure 4) is sent by the CN controller to the CN user plane (UPF) in the proposed architecture. There is no need for a session modification response message from the UPF. Hence, these reductions in the messages simplify both the session establishment and mobility procedures (to be discussed next). *Please note that even though RAN-User Plane (RAN-UP) and other RAN functions/messages are also necessary, we have shown only the CN functions in the call flow to keep the analysis tractable. However, keeping the RAN functions and the associated interactions out of the call flows is not likely to alter the conclusions drawn here. This note applies to mobility services also.*

4.2 User mobility

We consider user mobility as another service to illustrate the difference between the 5GS and the proposed architecture in terms of control flow and signalling exchange. Figure 5 and Figure 6 show the network entities, signalling and control message flow for the mobility service of the 5GS and proposed architecture, respectively. S-DU and T-DU represent source gNB-DU and target gNB-DU, respectively. Similarly, the Source-Centralized Unit-User Plane (S-CU-UP) and Target-Centralized Unit-User Plane (T-CU-UP) represent source gNB-CU-UP and target gNB-CU-UP, respectively. S-CU-CP and T-CU-CP represent source gNB-CU-CP and target gNB-CU-CP, respectively. Also, the interaction between the RAN controller and the CN controller is through the inter-controller interface, as shown in Figure 6.

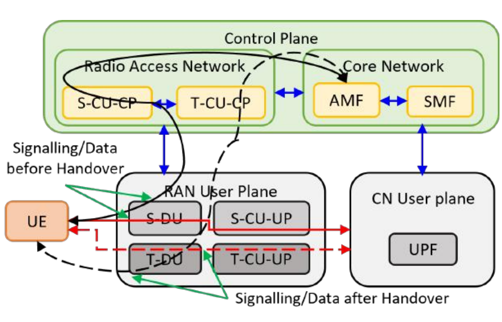

Fig. 5. Network entities, UE signalling and control message flow in case of mobility service for the 5GS architecture.

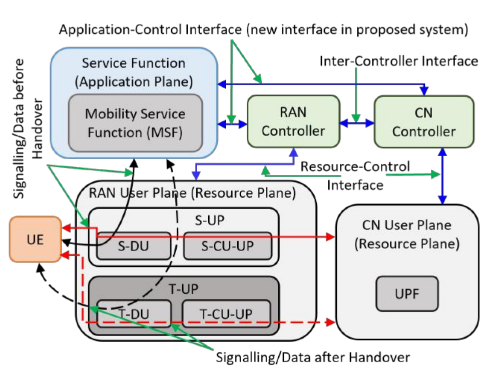

Fig. 6. Network entities, UE signalling and control message flow in case of mobility service for the proposed architecture.

Mobility call flow for the 5GS is available in [18]. Figure 7 here shows the mobility call flow which illustrates the mobility procedure of the proposed architecture. For the sake of simplicity, splitting S-UP into S-DU and S-CU-UP and T-UP into T-DU and T-CU-UP is not shown. However, the reason behind the simplification of mobility procedure/messages is the same as explained for PDU session establishment in Section 4.1.

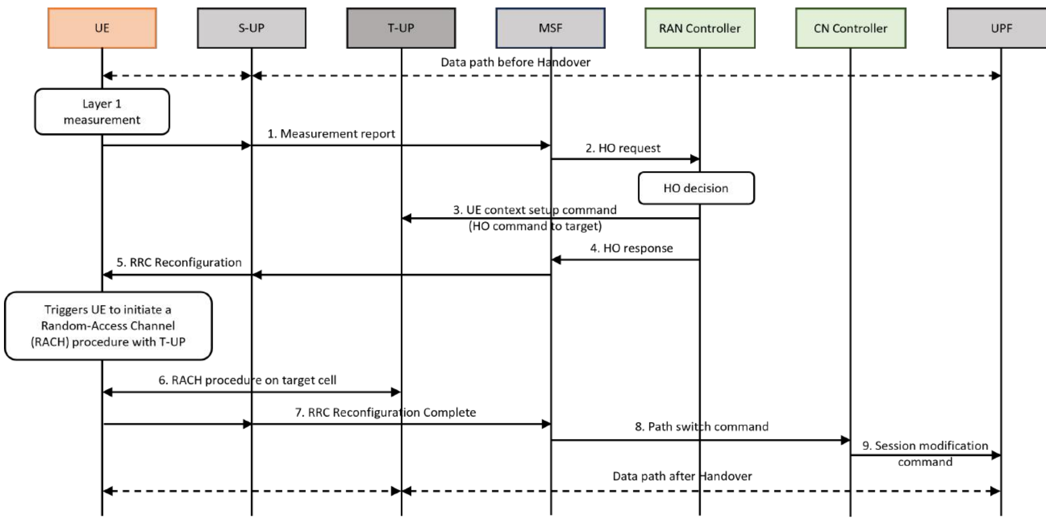

Fig. 7. Mobility procedure in the proposed architecture.

5 SYSTEM MODEL

This section presents the system model for the proposed architecture using PEPA. PEPA is a formal high-level language for the quantitative modelling of a distributed system [13]. Table 1a and Table 1b represent the system model for the proposed architecture for the PDU session establishment and mobility procedures, respectively. To explain the system models, we use the PDU session establishment (or session establishment) and mobility procedure (as shown in Figure 4 and Figure 7).

The session establishment procedure requires PSSF, CN controller and UPF as the key CN functions in the proposed architecture. These NFs are modelled as PEPA components. In addition, a UE is also modelled as a PEPA component. Each PEPA component (representing UE or a CN NF) goes through a set of states during the handling of the procedure. The individual component states are denoted by associating a unique number with the name of the component (e.g., $Pssf_1$, represents the first state of component, PSSF). Each component performs a set of actions, such as accessing the processor or sending a request/response. These actions are denoted in lowercase, and subscripts are added to provide further distinction (as $action_{actiondetail}$). For example, the notation for the action of PDU session establishment request and response can be req_{pduse} and rep_{pduse}, respectively. Each action is associated with a specific rate value, r. The rate (number of actions performed per unit time) models the expected duration of the action in the PEPA component and its values for different actions are taken as reference from [22], [23] and [24].

Let us now understand the details of modelling of NF states as shown in Table 1a. Consider UE as an example. The UE acquires the processor (acc_{uep}, r_a) in its initial state, Ue_1, and performs the processing action $(process, r_{iat})$ before sending a request. The second state, Ue_2, models the request (req_{pduse}, r_r) and response (rep_{pduse}, r_r) messages exchanged between UE and

PSSF for the PDU session establishment. NFs acquire processors to process a request/response. In Table 1a, UEP, PSSFP, CONP and UPFP are the processing entities for UE, PSSF, CN controller (CON) and UPF respectively. These processing entities are modelled such that each NF processor has two states. For instance, the first state of UEP, Uep_1, is for acquiring the processor (acc_{uep}), and the second state, Uep_2, performs the processing action ($process$). Similarly, the other NFs and their processing entities are modelled.

PDU session establishment	
PEPA Modules	**Code Description**
UE NF	$Ue_1 \stackrel{def}{=} (acc_{uep}, r_a).(process, r_{iat}).Ue_2$ $Ue_2 \stackrel{def}{=} (req_{pduse}, r_r).(rep_{pduse}, r_r).Ue_1$
PSSF NF	$Pssf_1 \stackrel{def}{=} (req_{pduse}, r_r).Pssf_2$ $Pssf_2 \stackrel{def}{=} (acc_{pssfp}, r_a).(process, r_p).Pssf_3$ $Pssf_3 \stackrel{def}{=} (req_{sc}, r_r).(rep_{sc}, r_r).Pssf_4$ $Pssf_4 \stackrel{def}{=} (acc_{pssfp}, r_a).(process, r_p).Pssf_5$ $Pssf_5 \stackrel{def}{=} (rep_{pduse}, r_r).Pssf_1$
CN Controller NF	$Con_1 \stackrel{def}{=} (req_{sc}, r_r).Con_2$ $Con_2 \stackrel{def}{=} (acc_{conp}, r_a).(process, r_p).Con_3$ $Con_3 \stackrel{def}{=} (req_{n4est}, r_r).(rep_{n4est}, r_r).Con_4$ $Con_4 \stackrel{def}{=} (acc_{conp}, r_a).(process, r_p).Con_5$ $Con_5 \stackrel{def}{=} (rep_{sc}, r_r).Con_1$
UPF NF	$Upf_1 \stackrel{def}{=} (req_{n4est}, r_r).Upf_2$ $Upf_2 \stackrel{def}{=} (acc_{upfp}, r_a).(process, r_p).Upf_1$
UE Processor	$Uep_1 \stackrel{def}{=} (acc_{uep}, r_a).Uep_2$ $Uep_2 \stackrel{def}{=} (process, r_p).Uep_1$
PSSF Processor	$Pssfp_2 \stackrel{def}{=} (process, r_p).Pssfp_1$
CN Controller Processor	$Conp_1 \stackrel{def}{=} (acc_{conp}, r_a).Conp_2$ $Conp_2 \stackrel{def}{=} (process, r_p).Conp_1$
UPF Processor	$Upfp_1 \stackrel{def}{=} (acc_{upfp}, r_a).Upfp_2$ $Upfp_2 \stackrel{def}{=} (process, r_p).Upfp_1$
System Equation	$(((Ue_1[n] \bowtie_{L_1} Pssf_1[N_{pssf}.N_{pssfp}.N_t])$ $\bowtie_{L_2} Con_1[N_{con}.N_{conp}.N_t])$ $\bowtie_{L_3} Upf_1[N_{upf}.N_{upfp}.N_t])$ $\bowtie_{L_4} (((Uep_1[n] \bowtie_{\phi} Pssfp_1[N_{pssf}.N_{pssfp}])$ $\bowtie_{\phi} Conp_1[N_{con}.N_{conp}])$ $\bowtie_{\phi} Upfp_1[N_{upf}.N_{upfp}])$
Cooperation Set	$L_1 = <req_{pduse}, rep_{pduse}>$ $L_2 = <req_{sc}, rep_{sc}>$ $L_3 = <req_{n4est}>$ $L_4 = <acc_{uep}, process, acc_{pssfp},$ $acc_{conp}, acc_{upfp}>$ $\phi = <>$

(a) PDU session establishment

Mobility	
PEPA Modules	**Code Description**
UE NF	$Ue_1 \stackrel{def}{=} (acc_{uep}, r_a).(measure, r_{iat}).Ue_2$ $Ue_2 \stackrel{def}{=} (reconfig, r_r).Ue_3$ $Ue_3 \stackrel{def}{=} (rachreq, r_r).(rachres, r_r).Ue_4$ $Ue_4 \stackrel{def}{=} (reconfigcomp, r_r).Ue_1$
T-UP NF	$Upt_1 \stackrel{def}{=} (pathsetup, r_r).Upt_2$ $Upt_2 \stackrel{def}{=} (acc_{uptp}, r_a).(process, r_p).Upt_3$ $Upt_3 \stackrel{def}{=} (rachreq, r_r).(rachres, r_r).Upt_1$
MSF NF	$Msf_1 \stackrel{def}{=} (measure, r_r).Msf_2$ $Msf_2 \stackrel{def}{=} (acc_{msfp}, r_a).(horeq, r_r).Msf_3$ $Msf_3 \stackrel{def}{=} (hores, r_r).Msf_4$ $Msf_4 \stackrel{def}{=} (acc_{msfp}, r_a).(reconfig, r_r).Msf_5$ $Msf_5 \stackrel{def}{=} (reconfigcomp, r_r).Msf_6$ $Msf_6 \stackrel{def}{=} (acc_{msfp}, r_a).$ $(pathswitch, r_r).Msf_1$
RAN Controller NF	$Ran_1 \stackrel{def}{=} (horeq, r_r).Ran_2$ $Ran_2 \stackrel{def}{=} (acc_{ranp}, r_a).(pathsetup, r_r)$ $.(hores, r_r).Ran_1$
CN Controller NF	$Cn_1 \stackrel{def}{=} (pathswitch, r_r).Cn_2$ $Cn_2 \stackrel{def}{=} (acc_{cnp}, r_a).(session, r_r).Cn_1$
UPF NF	$Upf_1 \stackrel{def}{=} (session, r_r).Upf_2$ $Upf_2 \stackrel{def}{=} (acc_{upfp}, r_a).(process, r_p).Upf_1$
UE Processor	$Uep_1 \stackrel{def}{=} (acc_{uep}, r_a).Uep_2$ $Uep_2 \stackrel{def}{=} (measure, r_{iat}).Uep_1$
T-UP Processor	$Uptp_1 \stackrel{def}{=} (acc_{uptp}, r_a).Uptp_2$ $Uptp_2 \stackrel{def}{=} (rachreq, r_r).Uptp1$ $+(rachres, r_r).Uptp_1$
MSF Processor	$Msfp_1 \stackrel{def}{=} (acc_{msfp}, r_a).Msfp_2$ $Msfp_2 \stackrel{def}{=} (horeq, r_r).Msfp_1+(reconfig, r_r)$ $.Msfp_1+(pathswitch, r_r).Msfp_1$
RAN Processor	$Ranp_1 \stackrel{def}{=} (acc_{ranp}, r_a).Ranp_2$ $Ranp_2 \stackrel{def}{=} (pathsetup, r_r).(hores, r_r).Ranp_1$
CN Processor	$Cnp_1 \stackrel{def}{=} (acc_{cnp}, r_a).Cnp_2$ $Cnp_2 \stackrel{def}{=} (session, r_r).Cnp_1$
UPF Processor	$Upfp_1 \stackrel{def}{=} (acc_{upfp}, r_a).Upfp_2$ $Upfp_2 \stackrel{def}{=} (session, r_r).Upfp_1$
System Equation	$(((((Ue_1[n] \bowtie_{L_1} Upt_1[N_{upt}.N_{uptp}.N_t])$ $\bowtie_{L_2} Msf1[N_{msf}.N_{msfp}.N_t])$ $\bowtie_{L_3} Ran_1[N_{ran}.N_{ranp}.N_t])$ $\bowtie_{L_4} Cn_1[N_{cn}.N_{cnp}.N_t])$ $\bowtie_{L_5} Upf_1[N_{upf}.N_{upfp}.N_t])$ $\bowtie_{L_6} (((((Uep_1[n] \bowtie_{\phi} Uptp_1[N_{upt}.N_{uptp}])$ $\bowtie_{\phi} Msfp_1[N_{msf}.N_{msfp}])$ $\bowtie_{\phi} Ranp_1[N_{ran}.N_{ranp}])$ $\bowtie_{\phi} Cnp_1[N_{cn}.N_{cnp}])$ $\bowtie_{\phi} Upfp_1[N_{upf}.N_{upfp}])$
Cooperation Set	$L_1 = < rachreq, rachres >$ $L_2 = <measure, reconfig,$ $reconfigcomp>$ $L_3 = < pathsetup, horeq, hores >$ $L_4 = < pathswitch >$ $L_5 = < session >$ $L_6 = <acc_{uep}, acc_{uptp}, acc_{msfp},$ $acc_{ranp}, acc_{cnp}, acc_{upfp}>$ $\phi = <>$

(b) Mobility

Table 1. System Model

As discussed in this section, the system model uses the following additional parameters: n denotes the number of UEs; N_{pssf}, N_{con}, and N_{upf} are the number of NF instances for PSSF, CN controller (CON), and UPF, respectively. Similarly, N_{pssfp}, N_{conp}, and N_{upfp} are the number of PSSF processors (PSSFPs), CN controller processors (CONPs) and UPF processors (UPFPs), respectively. Please note that each processor can handle a set of concurrent threads, N_t. Thus, the product $N_{nf} \cdot N_{nfp} \cdot N_t$ (where N_{nf} are the number of NFs, N_{nfp} are the number of processors for each NF as mentioned in the system model equation) represents the total number of threads for a type of NF. Moreover, the product $N_{nf} \cdot N_{nfp}$ is the total number of processors allocated to a type of NF, e.g., for UPF processor.

The system equation represents the overall system model. The cooperation operator ("$\bowtie$"), for example, A $\underset{L}{\bowtie}$ B, models the interactions between NFs A and B over the actions defined in the cooperation set L. It can be noted that it is possible that component A $\underset{L}{\bowtie}$ B will have different behaviour from component A $\underset{K}{\bowtie}$ B if L≠K. Let us consider an example from Figure 4, where PSSF and CN controller (CON) interact with each other for 'create session context request/response' req_{sc}/rep_{sc}. These actions are defined in cooperation set L_2, as shown in Table 1a. Therefore, the system equation $Pssf_1[N_{pssf}.N_{pssfp}.N_t] \underset{L_2}{\bowtie} Con_1[N_{con}.N_{conp}.N_t]$, models the interaction between PSSF and CN controller over the cooperation set L_2. In a similar way, the overall system equation, as shown in Table 1a and Table 1b represents the interaction between the various NFs as shown in the two call flows, Figure 4 and Figure 7, respectively.

6 PERFORMANCE EVALUATION

This section presents the performance comparison between the 5GS and the proposed architecture analysed using the PEPA Eclipse plug-in [25], a software tool integrated into the popular Eclipse platform. This tool supports various performance measures [23] as discussed below, which help evaluate the system performance. As mentioned before, the control plane performance has been evaluated here.

(1) **Session establishment rate** : The number of sessions established per unit time, measured for the action, rep_{pduse}, which describes the completion of the session establishment procedure. Similarly, to assess the performance of mobility service, the **number of successful handovers** is measured for the message session modification command (message 9 in Figure 7), represented as '$session$'(performed by UPF NF in Table 1b) signifying the completion of the mobility procedure.

(2) **Average response time**: It measures the UE waiting time for any specific request, e.g., 'session establishment' and reflects the system's operating speed. In this case, we consider the average response time as the duration required to complete the session establishment procedure. Similarly, we consider the mobility procedure's average response time as the duration to complete the mobility procedure.

(3) **Processor utilisation**: Processor utilisation measures the NFs processor capacity utilised during a procedure. The utilisation of any NF processor (for example, *PSSF* processor) while performing any procedure is derived from its population level analysis (one of the features available in the tool) [26].

(4) **Scalability**: Scalability *(S)*, in simple terms, measures a network's ability to increase or decrease its size, performance and cost in response to changes in system processing demands. Alternatively, according to Equation 1, scalability can be defined as the ratio between the productivity of a system at two configurations (configuration here implies the number of NFs used) having different scales, say m_1 and m_2 [27], which corresponds to the different numbers of NFs used in the network, say $m_1 = (1,1,1)$ and $m_2 = (3,3,1)$.

m_1 and m_2 details are discussed in subsection 6.1.1. The mathematical expression for scalability is given as [27]:

$$S(m_1, m_2) = \frac{C(m_2)}{C(m_1)}, \tag{1}$$

Where, $C(m)$ is the productivity of a system at the scale m, given by (Equation 2):

$$C(m) = \frac{t(m) \cdot r(m)}{U(m)}, \tag{2}$$

Where $t(m)$ is the average number of sessions established at scale m, $U(m)$ is the processor utilisation of the system (as defined in (3) of Section 6) at scale m, and $r(m)$ (Equation 3) is determined by evaluating the response time performance of the scaled system. We consider the following equation [27] to evaluate the performance function $r(m)$ by using the average response time $T(m)$, at scale m, with the target average response time T [23].

$$r(m) = \frac{1}{1 + T(m)/T}. \tag{3}$$

6.1 Results and Analysis

In this section, we present the performance results for 5GS and the proposed architecture in the case of PDU session establishment service and mobility service.

6.1.1 PDU Session Establishment Service. The performance analysis of the proposed architecture and the 5GS for the session establishment procedure is discussed in this section. Figure 8 and Figure 9 show the session establishment rate with respect to the number of UEs for 5GS and the proposed architecture using two different configurations. For instance, $(N_{pssf}, N_{con}, N_{upf}) = (1,1,1)$ for the proposed architecture is the basic configuration (m_1) with single NF instances assigned to each NF, i.e., to PSSF, CON, UPF and $(N_{pssf}, N_{con}, N_{upf}) = (3,3,1)$ is the configuration for a scaled system (m_2) with three NF instances assigned to PSSF and CON while one to UPF. Similarly, basic and the scaled configuration for 5GS is defined as $(N_{amf}, N_{smf}, N_{upf}) = (1,1,1)$ and $(N_{amf}, N_{smf}, N_{upf}) = (3,3,1)$, respectively.

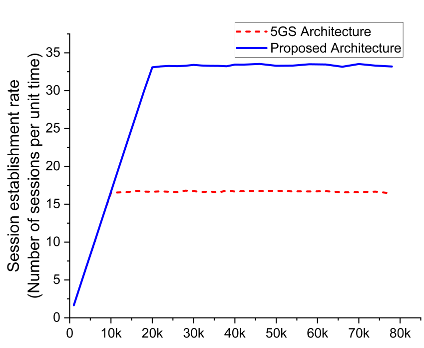

Fig. 8. Session establishment rate for the proposed and the 5GS architecture with the basic configuration (m_1 = (1,1,1)).

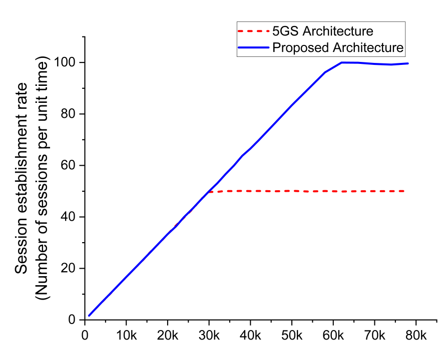

Fig. 9. Session establishment rate for the proposed and the 5GS architecture with the scaled configuration (m_2 = (3,3,1)).

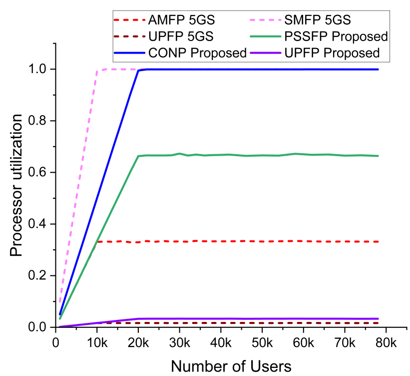

Fig. 10. Processor utilisation of session establishment for the proposed and the 5GS architecture with the basic configuration (m_1 = (1,1,1)).

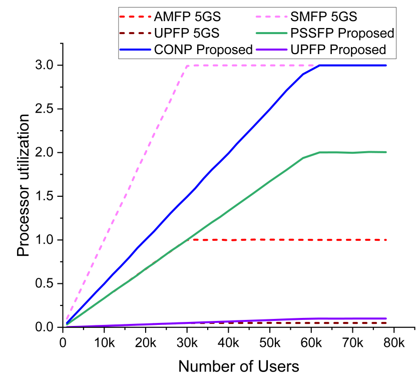

Fig. 11. Processor utilisation of session establishment for the proposed and the 5GS architecture with scaled configuration (m_2 = (3,3,1)).

Results show that the proposed architecture can achieve a higher session establishment rate compared to the 5GS in case of both basic and scaled configurations. Although the session establishment rate has increased using a scaled configuration for both the proposed and the 5GS architectures compared to the session establishment rate achieved using a basic configuration, the proposed architecture achieves a higher session establishment rate than the 5GS. The saturation point for 5GS, as shown in Figure 8, is around 10,000 users i.e. it can serve a maximum number of 10,000 users in case of basic configuration, while the session establishment rate for the proposed architecture saturates at around 20,000 users. Similarly, Figure 9 shows that 5GS saturates at around 34,000 users in scaled configuration whereas the proposed architecture saturates at 62,000 users. As the saturation point is reached, the network starts dropping the incoming requests from users. This means that with the given number of processors/NFs, the proposed architecture can achieve a higher session establishment rate. The processor utilisation for all NFs of the 5GS and the proposed architecture for basic and the scaled configuration are shown in Figure 10 and Figure 11, respectively. It should be observed that the saturation point for processor utilisation is much higher for the proposed architecture viz-a-viz the 5GS. For instance, the PSSFP reaches its maximum utilisation explaining the saturation point for the session establishment rate. However, at this point, CONP and UPFP are not fully utilised. These results show that the request processing chain fails if an NF becomes a bottleneck for the consecutive chain.

Scalability for the 5GS and the proposed architecture is evaluated as per Equation 1. It is plotted in Figure 12 on the basis of the results obtained for session establishment rate, average response time and utilisation from the PEPA-based simulation and modelling. As stated earlier, we consider the two configurations m_1 (basic configuration with single NF instances assigned to each NF, i.e., to PSSF, CON, UPF and represented as $(N_{pssf}, N_{con}, N_{upf}) = (1,1,1)$) and m_2 (scaled configuration with three NF instances assigned to PSSF and CON while one to UPF and represented as $(N_{pssf}, N_{con}, N_{upf}) = (3,3,1)$) for estimating the scalability metric. Figure 12 shows that the 5GS can serve 10,000 users for a basic configuration, and the proposed architecture can serve 20,000 users. Similarly, the 5GS reaches its saturation point at 34,000 users, and the proposed architecture saturates at 62,000 users for scaled configuration. As a result, the curve emphasizes that the proposed architecture has the capacity to support a larger number of users, reaching a saturation point later than that of the 5GS. Besides, the proposed architecture is more

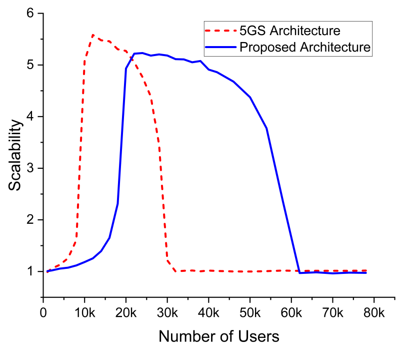

Fig. 12. Scalability in case of PDU session establishment for the proposed and the 5GS architecture.

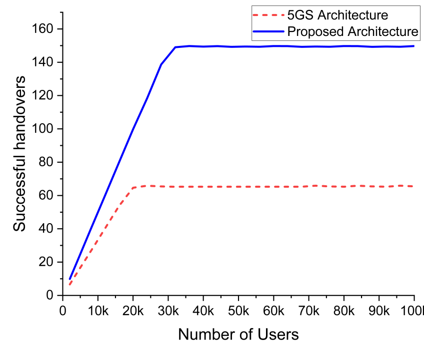

Fig. 13. Number of successful handovers per unit time for the proposed and the 5GS architecture with the basic configuration (m_1 = (1,1,1,1,1,1)).

scalable with increased users for the same number of NFs/processors. Please note that a similar explanation for all the performance measures (successful handovers, processor utilisation and scalability) holds in the case of mobility service.

6.1.2 Mobility Service. This section presents the comparative analysis of the 5GS and the proposed architecture for the mobility service. Similar to the session establishment, the analysis is performed for the basic and the scaled configurations. The basic configuration for the proposed architecture is given as $(N_{upt}, N_{msf}, N_{ran}, N_{cn}, N_{upf}) = (1,2,2,1,1)$ and for the 5GS architecture is $(N_{sdu}, N_{scu}, N_{tdu}, N_{tcu}, N_{amf}, N_{smf}, N_{upf}) = (1,1,1,1,1,1,1)$. Similarly, the scaled configuration for the proposed architecture is $(N_{upt}, N_{msf}, N_{ran}, N_{cn}, N_{upf}) = (3,6,6,3,3)$ and for the 5GS architecture is given as $(N_{sdu}, N_{scu}, N_{tdu}, N_{tcu}, N_{amf}, N_{smf}, N_{upf}) = (3,3,3,3,3,3,3)$. Here N_{upt}, N_{msf}, N_{ran}, N_{cn}, N_{upf} are the number of Target-User Plane (T-UP), MSF, RAN controller, CN controller and UPF respectively in the system model. Similarly, N_{sdu}, N_{scu}, N_{tdu}, N_{tcu}, N_{amf}, N_{smf}, N_{upf} are the number of S-DU, S-CU, T-DU, T-CU, AMF, SMF, and UPF respectively. Please note that for brevity, we have not split S-CU into S-CU-CP and S-CU-UP and T-CU into T-CU-CP and T-CU-UP while modelling the mobility call flow procedure for the 5GS. Further, we use an equal number of functions and associated processors to the 5GS and the proposed architecture for justified comparison.

Figure 13 and Figure 14 show that the proposed architecture serves more successful handovers per unit time compared to the 5GS for both the basic and the scaled configurations, respectively. The saturation point for the 5GS is 20,000 users, while for the proposed architecture, the saturation point is 30,000 users for the basic configuration. Similarly, in the scaled configuration, the saturation point for the 5GS is around 60,000 users, while for the proposed, the saturation is around 90,000 users. The number of successful handovers per unit of time has increased using a scaled configuration for both architectures.

Figure 15 and Figure 16 are the results of processor utilisation for both the 5GS and the proposed architecture. A similar observation is noted here as well, indicating that the saturation point for processor utilisation is significantly higher for the proposed architecture viz-a-viz the 5GS. As an illustration, the DPTP reaches its maximum utilisation, elucidating the saturation point for the number of successful handovers per unit time. At this point, other processors are not fully utilised. These findings draw a similar conclusion that the request processing chain fails

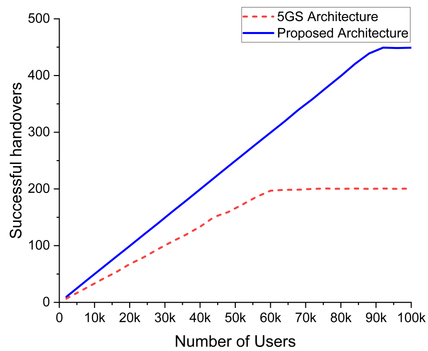

Fig. 14. Number of successful handovers per unit time for the proposed and the 5GS architecture with the scaled configuration (m_2 = (3,3,3,3,3,3,3)).

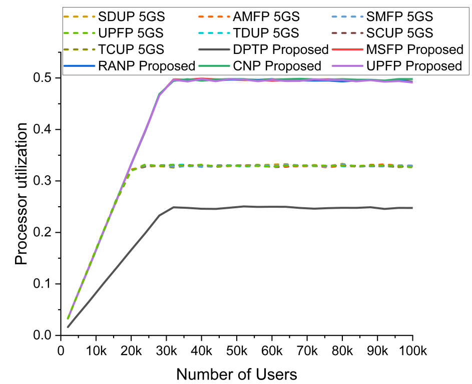

Fig. 15. Processor utilisation in case of mobility service for the proposed and the 5GS architecture with the basic configuration (m_1 = (1,1,1,1,1,1,1)).

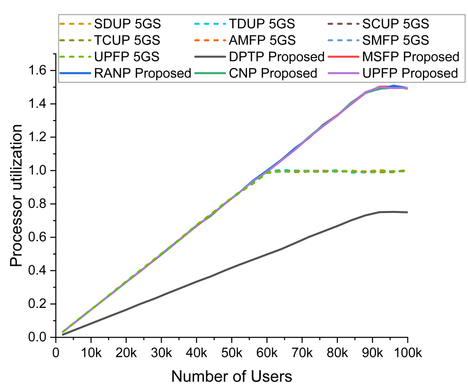

Fig. 16. Processor utilisation in case of mobility service for the proposed and the 5GS architecture with the scaled configuration (m_2 = (3,3,3,3,3,3,3)).

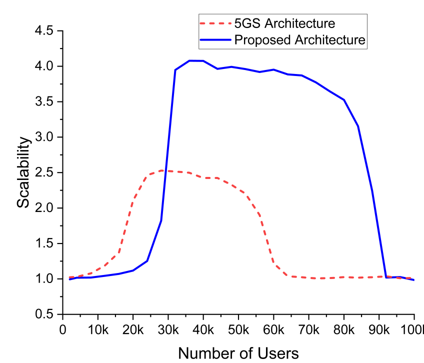

Fig. 17. Scalability in case of mobility service for the proposed and the 5GS architecture.

if an NF becomes a bottleneck in the consecutive chain. Figure 17 shows the scalability results in the case of mobility service for 5GS and the proposed architectures. It can be observed from the scalability results that 5GS reaches its saturation point earlier than the proposed architecture and the proposed architecture is more scalable.

As highlighted in Section 3, there is no change on UE both with respect to signalling and data transfer in the proposed architecture viz-a-viz the 5GS; the performance gain in the proposed system is only due to the reorganization of the functionality on the network side and simplification in protocols between network functions.

7 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel mobile network architecture to separate the handling of UE signalling from the user plane control (resource control) functionality, enhancing the modularity, scalability, and flexibility of the network control plane. The transposition of UE signalling handling from control plane to service plane is a paradigm shift. It leads to simplified protocols and opens up new ways to implement use case specific signalling in mobile networks. The proposed architecture also has improved alignment with SDN and the principle of separation of concern. We have considered PDU session establishment and mobility services as examples to analyse the performance of the proposed architecture using the PEPA-based simulation method. Based on the performance results and other benefits, it can be concluded that the proposed architecture is a promising option for future networks to handle vast and diverse traffic demands and a much larger number of users. We plan to extend this work to analyse other features/services of mobile networks, such as network slicing, protocol design between (signalling) service functions and the control plane, and addressing security threats in mobile networks in future.

REFERENCES

[1] FG NET-2030. Achievements of ITU-T Focus Group on Network 2030. *7th SG13 Regional Workshop for Africa on ITU-T Standardization Work on Future Networks: Towards a Better Future for Africa (Abuja, Nigeria)*, 2020.

[2] SWG IMT-2030. Framework and overall objectives of the future development of IMT for 2030 and beyond. *Technical Recommendation*, 2023.

[3] Shah, Rinku and Kumar, Vikas and Vutukuru, Mythili and Kulkarni, Purushottam. TurboEPC: Leveraging Data plane Programmability to Accelerate the Mobile Packet Core. *ACM SIGCOMM Symposium on SDN Research (SOSR)*, pages 1–13, 2020.

[4] Taksande, Pradnya Kiri and Jha, Pranav and Karandikar, Abhay. Dual Connectivity Support in 5G Networks: An SDN based approach. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2019.

[5] Akshatha, Nayak M. and Jha, Pranav and Karandikar, Abhay. A Centralized SDN Architecture for the 5G Cellular Network. In *IEEE 5G World Forum (5GWF)*, pages 147–152, 2018.

[6] Roozbeh, Amir. Distributed Cloud and De-centralized Control Plane: A Proposal for Scalable Control Plane for 5G. In *IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*, pages 348–353, 2015.

[7] Ueda, Tetsuro and Idoue, Akira and Utsunomiya, Eiji. Hierarchical and Distributed Software-Defined Network to Reduce Control Load. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2019.

[8] Abdulghaffar, Abdulaziz and Mahmoud, Ashraf and Abu-Amara, Marwan and Sheltami, Tarek. Modeling and Evaluation of Software Defined Networking Based 5G Core Network Architecture. *IEEE Access*, 9:10179–10198, 2021.

[9] Zuo, Qingyun and Chen, Ming and Ding, Ke and Xu, Bo. On Generality of the Data Plane and Scalability of the Control Plane in Software-Defined Networking. *China Communications*, 11(2):55–64, 2014.

[10] Hu, Jie and Lin, Chuang and Li, Xiangyang and Huang, Jiwei. Scalability of Control Planes for Software Defined Networks: Modeling and Evaluation. In *IEEE 22nd International Symposium of Quality of Service (IWQoS)*, pages 147–152, 2014.

[11] Amin Tootoonchian and Yashar Ganjali. Hyperflow: A distributed control plane for openflow. In *Proceedings of the Internet Network Management Conference on Research on Enterprise Networking*, page 3, 2010.

[12] Jordan Raychev, Diyana Kinaneva, Georgi Hristov, and Plamen Zahariev. Optimizing sdn control plane scalability by efficient switch to controller migration. In *27th National Conference with International Participation (TELECOM)*, pages 42–45, 2019.

[13] Hillston, J. A Compositional Approach to Performance Modelling. *Distinguished Dissertations in Computer Science. Cambridge University Press.*, 1996.

[14] 3GPP TS 23.501. System Architecture for the 5G System. *Technical Specification*, 2018.

[15] ITU-T Y.3300. Framework of Software-Defined Networking. *Recommendation*, 2014.

[16] ONF TR-502. SDN Architecture. *Technical Specification*, Issue 1, 2014.

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

[18] 3GPP TS 23.502. Procedures for the 5G System. *Technical Specification*, 2018.

[19] 3GPP TS 33.512. 5G; 5G Security Assurance Specification (SCAS); Access and Mobility Management Function (AMF) . *Technical Specification*, 2022.

[20] 3GPP TS 33.511. 5G; Security Assurance Specification (SCAS) for the next generation Node B (gNodeB) network product class . *Technical Specification*, 2022.

[21] ENISA. SECURITY IN 5G SPECIFICATIONS; Controls in 3GPP Security Specifications (5G SA). *Technical Specification*, 2021.

[22] Oliveira, E. M. R and Viana, A. C. and Naveen, K. P. and Sarraute, C. Mobile data traffic modeling: Revealing temporal facets. *Computer Networks*, 112:176–193, 2017.
[23] Arteaga, C. H. T. and Ordoñez, A. and Rendon, O. M. C. Scalability and Performance Analysis in 5G Core Network Slicing. *IEEE Access*, 8:142086–142100, 2020.
[24] Arteaga, C. H. T. A 5G Core Network Prototype. *Computer Networks*, 112:176–193, 2019.
[25] Tribastone, Mirco and Duguid, Adam and Gilmore, Stephen. The PEPA Eclipse Plugin. *SIGMETRICS Perform. Eval. Rev.*, 36(4):28–33, 2009.
[26] B.F. Nielsen, F Nielson, H Pilegaard, M.J.A. Smith, E. Yüksel, K. Zeng, and L. Zhang. Roadmap document on stochastic analysis. *ArXiv*, abs/1209.6578, 2012.
[27] Jogalekar, P. and Woodside, M. Evaluating the scalability of distributed systems. *IEEE Transactions on Parallel and Distributed Systems*, 11(6):589–603, 2000.