A Dynamic Pricing and Leasing Module for 5G Networks

Kostas Chounos, Apostolos Apostolaras and Thanasis Korakis
Dept. of Electrical and Computer Engineering, University of Thessaly, and CERTH, Greece

Abstract—In this paper, a novel framework for enabling dynamic Policy and Charging Rules in modern 5G multi-tenancy environments is designed and presented. More specifically, adaptive pricing schemes are utilized by MNOs for instantiating tailored slices, when the resources of a Mobile Network Operator (MNO) are insufficient to serve the connected users. A scheme where an MNO seeks and leases additional resources from an Infrastructure Provider (InP) is examined in this paper. Their negotiation is modeled as a Service Level Agreement (SLA) and the dynamic charging rules are implemented as an extension of the 5G Policy Charging Function (PCF). The analytical structure of the proposed architecture as well as all the methods and signals developed for MNO and InP collaboration, are described in the context of this work. Finally, extensive testbed experimentation proves proper functioning of the proposed framework, under both UDP and TCP additional traffic demands.

Index Terms—5G, Dynamic Charging Rules, Tailored Slicing, Pricing, Resource Leasing

I. INTRODUCTION

Over the last years, we are witnessing an indisputable rise on the users’ demand for consuming multimedia data traffic. This situation exerts pressure on network operators for fulfilling the end users requirements. Currently, the infrastructure-centric approach of the past network architectures is proved inadequate to handle the increasing traffic volume and impacts negatively the operational expenditures of the operators. As a result, telecommunications service providers and mobile network operators have been seeking for alternative non-costly and performance-efficient solutions that would be able to accommodate the ever-increasing users’ demand.

Unavoidably, the transition from large scale base stations to small remote radio heads supporting disaggregation of the protocol stack to enable agile deployments, is essential to reconsider the cellular network architecture. Agile network deployments supporting low latency and higher throughput communications with lower maintenance and operational costs are currently required for enabling ultra reliable and ubiquitous communication services. Towards this pace, 5G is anticipated to change drastically the way that the information is being used, processed and handled, supporting smarter and higher performing flexible networks.

Recent technological advances like Network Functions Virtualization (NFV) [1], [2], [3] and Software Defined Networks (SDN) [4], [5], [6] as well as ongoing standardization efforts [7], [8] render the new 5G architecture highly attractive by encompassing NFV, SDN and cloud technologies, which offer diversified capabilities regarding the programmability of the network functionalities. Those technologies have introduced the following principles on the network architecture design. Namely:

- **Disaggregation:** It is the procedure of the vertical and distinct separation of integrated systems into independent components being equipped with open interfaces. The adoption of network function disaggregation principles, enables network operators to select the best combination among several individual hardware and software components while eliminating integrated systems and protocols lock. All the networking software needed to enable networking functionality is broken down and thus can be operated more efficiently, which leads to improved physical resources utilization.

- **Virtualization:** It provides the ability to instantiate, run and operate multiple independent copies/replicas of the network functionalities on a common hardware platform. From a network perspective, the concept of slicing [1] has been recently introduced in 5G architecture to enable the partitioning of the underlying network infrastructure in independent logical networks and guarantee the isolation among network tenants.

- **Commodification:** of information data and data transactions combined with the provision of inexpensive and high performance infrastructure (both virtual and physical) and spectrum availability. It is the ability of elastic scaling on the virtual network components across commodity hardware bricks as workload and networking requirements dictate.

To this end, the rise of 5G technology is expected to create new collaboration opportunities among operators and enable new Business to Business (B2B) models [9], [10], [11]. Existing stakeholders that used to consume data and infrastructure services will update their role and will share their resources acting as potential infrastructure providers (InPs) in the upcoming 5G network models and under the introduction of network slicing. Thus, there is an immediate need for deploying proper resource brokering mechanisms to the slice providers, as significant economic benefits can occur. Additionally, several challenges on how the slice providers will economically interact with verticals and resource providers are being emerged in 5G networks.

In this work, we design a novel framework for virtual MNOs\(^1\) which enables dynamic allocation of resources to expand their network’s capacity through resource leasing and

\(^1\)With the term MNO, we refer also to mobile network operators that offer their services virtually.
by negotiating an SLA agreement with the InP. Thus, MNOs in need of additional resources, due to a possible temporal or constant capacity crunch can request on demand from other MNOs or InPs resources to fulfill their needs and provide sufficient services to users guaranteeing performance. We conduct extensive testbed experimentation to thoroughly assess our design. Relying on our recent theoretical work [12] that models the interactions among users, MNOs and infrastructure providers in a 5G network context, and utilizing the designated optimal policies for pricing and leasing of resources, we design the blueprints for a Policy and Charging Control (PCC) core network entity that it can dynamically coordinate pricing and leasing decisions. The key contributions of this work can be summarized as follows:

- **Dynamic PCC Rules:** We design a novel framework that allows dynamic leasing and pricing of networking resources. Our solution monitors the users demands and assist operators to apply optimal pricing strategies. Besides, in the case of network capacity depletion, virtual MNOs are assisted by our solution to acquire the appropriate amount of additional resources through leasing from the InP in order to accommodate the extra demand.

- **Extensive evaluation:** We assess the proposed solution using real testbed experimentation utilizing OpenAirInterface [13] and FlexRAN [14] software. Leveraging our theoretical pricing and leasing framework[12], we demonstrate the applicability of our framework in real 5G systems. Extensive performance testing has been conducted in order to validate and assess the proposed framework’s proper functionality.

The rest of this paper is organized as follows: In Section II the related work is presented. Furthermore, in Section III the architecture of the proposed framework is described analytically. Section IV presents the evaluation scenarios as well as the collected experimental results. Finally, Section V concludes the paper and presents possible extensions.

II. RELATED WORK

**Standardization Efforts:** Three different Specification Groups (SG) were formed by 3GPP, the Radio Access Network (TSG-RAN), the Services & System Aspects (TSG-SA) and the Core Network & Terminals (TSG-CT). These groups are responsible for creating holistic architectural designs for the upcoming network generations. Huge emphasis is being lately given to the softwarization and virtualization according to which, all the software needed for the base stations is ported on the cloud. The new network architecture leverages Virtual Network Functions (VNFs) to compose the core network entities, as a way of implementing modern architectures under a more accessible, distributed and flexible way. Consequently, the symmetrical use of common physical infrastructures for more than one competitive MNOs/InPs and under the use of Service Based Architectures (SBAs), are now in the spotlight of the research approaches. However, the simultaneous use of common physical resources raises several challenges as well. Distinct separation as well as proper orchestration and resource management, are now essential for achieving smooth operation at 5G networks. Network slicing offers huge opportunities for efficient utilization of the limited network resources. The 3GPP (TSG-SA) working group is responsible to capture in detail, how network slices will be integrated at 5G core architectures. More specifically, the Technical Specifications (TS) 23.501[7], 23.502[8] and 23.503[15] contain all the information regarding System Architecture, 5G System and PCF aspects accordingly.

**Research Efforts:** With emphasis at some of the most vital details for networks slices, a high-level view can be defined as follows. Network slice is an independent end-to-end logical network which runs on a shared physical infrastructure and is capable of providing an agreed quality of service. This type of agreement is established between two entities (e.g. InP and MNO) and it is referred as an SLA. As the core network will be implemented in a cloud and generally all architectures are now software oriented, multiple approaches for network slicing are given in the literature so far [1]. Moreover, Oladejo et al. in [16] describe practical considerations for how multi-tenancy network slicing, can affect the overall network capacity. Specifically, they examine how the number of users in a slice, the total number of slices and the transmit power affect the capacity of an V-MNO. In addition to the tailored performance, each slice will have the flexibility to operate either under unified or separate set of NFs, as shown in Fig.1. In such way, the ability for discriminated pricing at each slice, will be enabled at 5G networks. Appropriate techniques for dynamic pricing of the leased resources are being investigated and tested in the context of 5G networks [17], [18].

Extensive functionality for policy based control of traffic and user services, is now essential for the upcoming network architectures. By replacing the former 4G Policy Control and Charging Rules function (PCRF), the PCF will undertake this critical role. Therefore, transparent control for monitoring the consumption of network resources (slices) will be provided through PCF. Both online (OCS) and offline (OFCS) charging are now implemented as an entity called Convergent Charging System (CCS) and enclosed in 5G Charging Function (CHF). Based on the 3GPP PCF [15], the Policy and Charging control (PCC) rules incorporate all the information for enabling the user plane detection of the policy control and proper charging for a service data flow. Additionally, the tight coordination of PCF with different fundamental 5G Network Functions...
(NFs) like UDR, AMF, SMF and CHF, enables additional policy control functionalities besides charging. Indicatively, rules for how the users and data sessions are controlled in terms of Quality of Service (QoS), Geographical Restrictions and available RAN selection are some of the extended PCF responsibilities.

III. PROPOSED FRAMEWORK

In this work, we design and present a framework which extends the standard operation of the PCF function. During real-time service delivery, PCF is in charge to coordinate and enforce control and user plane rules. The intellectual merit of this work develops a charging and pricing solution that helps MNOs that face capacity shortcomings and their service capability does not suffice due to the increased users' demand. The proposed design is used to monitor the control and data plane traffic and apply specific rules that assist MNOs to acquire additional resources through leasing from the InP or from nearby MNOs when the MNO faces service insufficiency as well as to apply optimal pricing to its users. Therefore, the MNO in need and the InP interact with each other and transact for leasing resources. The requested resources after being purchased by the MNO, are attached to its network slice, thus extending its service capability with extra capacity to serve its users. The analytical call flow for the proposed architecture as well as the interactions between the MNO and the InP are illustrated in Fig. 2.

A. Overview of the Leasing Interaction

In the context of this work, we leverage on our theoretical framework [12], which considers the problem for MNOs of leasing resources, servicing and pricing mobile users. Through a three-stage Stackelberg game solved by backward induction, the optimal pricing decisions when MNO’s supplying capacity suffices users’ demand, as well as when supplying capacity is deficient are calculated. In the latter case, when MNOs supplying capacity doesn’t suffice, the MNO requests through leasing the proper amount of additional resources from the InP that not only resolves its deficiency but also incurs the maximum benefit given the leasing cost by the InP. The interactions among the UEs, the MNO and the InP can be summarized as follows: The operation of the proposed solution is performed real-time.

The MNO, firstly assesses the total demand that is requested by the associated users and depending on that, MNO checks his capacity availability and determines the amount of resources that needs (or not) to purchase from the infrastructure provider in order to satisfy this demand. In sequence, the MNO determines the optimal pricing that increases its profits. Thus, the MNO announces that price to the users and then, the users by taking into consideration the price that the MNO had already announced, demand on their side the corresponding rate that maximizes their own benefit, as this is expressed by their payoff function. For more details, on the modeling of the game among the users, the MNO and the InP, the interested reader is referred to [12].

B. Call Flow

Fig. 2 shows an illustration of the call flow that is initiated when the MNOs network controller detects service performance insufficiency and specifies the need for the MNO to acquire additional resources. All the modules and signaling that is being communicated in the proposed framework are described below. The implementation of the proposed framework was developed by using Python programming language. Furthermore, the communication among the network entities (e.g. UEs, MNO and InP) is performed through socket messaging and by utilizing the corresponding Python library.

• MNO Resource Availability Check: This module is used by the MNO to assess the capacity of its network resources (“gNB1”) whether they are capable for providing sufficient services to its connected UEs. If the resources of MNO do not suffice, a signal is used for requesting additional capacity given the leasing cost price $P_{\text{InP}}$ that the InP imposes to the MNO.

• $P_{\text{MNO}}$ Calculation: This module receives all the necessary inputs for determining the optimal price $P_{\text{MNO}}$, that the MNO announces to its users.

• $P_{\text{MNO}}$ Announcement: This signal is sent between the MNO and UEs and it refers to the announcement to the UEs for the price $P_{\text{MNO}}$.

• UE Demands Request: In response to the pricing $P_{\text{MNO}}$ Announcement signal, this signal is used to communicate to the MNO the respective UE demand.

• Request for Additional Resources-Capacity given the leasing cost $P_{\text{InP}}$: This signal is used by the MNO to get notified for the leasing cost $P_{\text{InP}}$ set by the InP and to request the amount of capacity that wishes to lease from the InP given the cost $P_{\text{InP}}$.

• $P_{\text{InP}}$ Calculation: This module is used by the InP to determine the leasing cost $P_{\text{InP}}$ to be announced to the MNO.

• InP Resource Availability Check: Similarly with MNO, the InP should assess its resource availability. The provisioned resources could be either cellular (e.g. from the slice that corresponds to the “gNB2” cell) or WiFi network resources. Part of those resources can be attached to the MNO’s slice (e.g “gNB1”) upon request to serve the increased users’ demand.

• Normalization of the demand volume to a percentage of utilized resources: In this module and after successfully accepting the capacity request from the MNO, the InP instantiates the procedure to provide the MNO with the requested amount of capacity that corresponds to the virtual resources to be attached to the MNO’s slice. It should be noted that the slice reconfiguration with the additional amount of resources is achieved through FlexRAN [14], a programmable platform for managing real-time and dynamically the physical resources of a 5G network. More specifically, leveraging OpenAirInterface [13] Core Network and Radio Access Network (OAI-CN/OAI-RAN), FlexRAN can provide flexible re-configuration and scalability of network
slices in radio access networks (RANs). Furthermore, the description and the transmission of the requests/responses between the Master Controller and the RAN Agents is performed through the FlexRAN API that utilizes curl command and json files. However in the slice creation, FlexRAN takes as an input the percentage that the new slice will occupy in the overall’s cell capacity. Based on that, we convert the requested demand from the MNO (Mbps) to cell percentage. An analytical description for the cell percentage calculation, is given in Section IV.

- **Slice Instantiation/Re-configuration**: Finally, after retrieving the downlink/uplink percentages from the Demand Conversion function, we communicate through the appropriate json descriptions to the target gNB’s FlexRAN Agent\(^2\) in order to instantiate or re-configure the new slice.

- **Response with \(P_{\text{InP}}\) and Slice Status**: This signal is responsible either to inform the MNO for the \(P_{\text{InP}}\) or to report the status for a requested slice.

- **PCC Rule Creation**: In this module and after being informed for the cost \(P_{\text{InP}}\), the MNO undertakes to create a custom PCC rule for the UE which requested additional resources (leased slice). As mentioned in a prior section, a PCC rule [15] is a set of information which enables the detection of a service data flow and provides parameters for proper policy and/or charging controls. Each PCC rule functions under a set of operations in the 5G networks. Indicatively, the PCF function can activate, modify and deactivate a PCC rule at any time. The PCC rules are divided into two types, the predefined and the dynamic. The former are predefined rules which are configured into SMF and only referenced by the PCF, while the latter can be dynamically provisioned by the PCF to the SMF. It is worth to be noted that modifications are applicable only to dynamic PCC rules. Additionally, the session management related policy control functionality of the PCF in combination with the PCC rules, enable the functions for dynamic policy and charging control as well as event reporting for service data flows in 5G systems. In the context of this work and through the use of dynamic PCC rules, the credit info updated dynamically for each UE in a per slice basis and based on the announced price cost \(P_{\text{InP}}\). More specifically, proper PCF decisions are taken and provided to the SMF for each leased slice. These decisions consist of both PCC rules as well as with the appropriate customized PDU session related attributes.

IV. EVALUATION

In this Section, we evaluate the performance of the proposed framework and we analyze the collected results. We conducted extensive experimentation using real hardware equipment in the NITOS Testbed (University of Thessaly, Greece.) [19]. There, a wide variety of software and hardware supporting heterogeneous communication technologies are available to the experimenter 24/7.

A. Topology and Configurations

For the experimentation scenarios under consideration, we used 6 testbed nodes in the network topology shown in Fig. 3. Specifically, 4 of the nodes equipped with Huawei’s E-392 [20] LTE usb sticks were instantiated as user equipments (UEs). Moreover, we used one node to act as host for the core network for which we utilised the OpenAirInterface
framework [13] (OAI-CN) to instantiate it. For the radio access network we used the OAI-RAN and the USRP B210 [21] Software Defined Radio equipment. The gNB is configured with a 20 MHz channel bandwidth, 64 QAM modulation, single input - single output (SISO) antenna and employs a Frequency Division Duplexing (FDD) channel access scheme. To properly configure the capacity of the MNO’s RAN slice we used the FlexRAN framework which takes as an input configuration parameter the percentage of the network capacity that will be allocated to the slice. Relying on a pre-assessment of the total provisioned capacity of the radio infrastructure, we measured the maximum capacity (throughput rate in bps) to be 72.2 Mbps for UDP traffic and 47.5 Mbps for TCP respectively, while the theoretical maximum rate is 75 Mbps (according to the base station RF configuration).

B. Experimental Scenarios

1) UDP Traffic Evaluation: The aim of this scenario is to showcase the adaptability of the proposed framework to the network changes. The MNO has been allocated a slice of resources (part of the gNB’s capacity) that are initially owned by the InP. The slice settings can be re-configured on-the-fly and on demand according to the MNO’s needs to service its users by using the FlexRAN API. Initially, the MNO’s slice is configured with a pre-defined bandwidth budget that corresponds to the 16 % of the gNB’s 72.25 Mbps maximum capacity. (Therefore the MNO’s initial capacity is approximately 12 Mbps). We used 4 testbed nodes to act as UEs and we initiated multiple UDP traffic flows (5 per UE) requesting download services using the iperf command. Figure 4a shows the respective collected measurements. In the first 30 seconds of the experiment, the MNO’s supplying capacity (12 Mbps) is less than the users’ demand \( r = 13 \) Mbps and users receive a total service of around 12 Mbps. Therefore, there is an insufficiency gap for servicing the full rate of users demand. Afterwards, between 31 – 60 secs, the MNO announces the pricing to its users \( p = 0.47 \) for each video segment and the users respond by requesting an extra demand based on that price\(^3\). As a result the total user demand increases and exceeds the initial MNO’s servicing capability. Immediately, to mitigate this incident, the MNO determines its deficit and seeks to acquire additional resources from the InP. The InP leases the additional resources and the MNO purchases them at a cost \( P_{\text{InP}} = 0.4 \$ \). In this period the MNO’s slice is reconfigured accordingly using FlexRAN to support this demand. Next, between 61 – 90 secs, arrivals of new users occur and the total demand increases. The proposed framework monitors the increase in this demand and the MNO’s slice is reconfigured accordingly.

2) MPEG-DASH Video Streaming: In this scenario, we assess the performance of the proposed framework under Dynamic Adaptive Streaming over HTTP (MPEG-DASH) [22]. An APACHE HTTP web server [23] was placed in the host node of the core network to serve the requested video streams from the UEs. Each DASH formatted video stream, consists of chunked videos with multiple available stream-quality levels. These video parts are also called segments, which in the examined case have a time duration of 1 sec. Before servicing a video segment to the client, the MPEG-DASH algorithm selects the appropriate video quality that can be supported based on some specific metrics. The key metric that we utilized in our setup refers to the experienced rate. Therefore, the MPEG-DASH selects the maximum quality that can be achieved for each segment given the link’s capacity. The minimum segment quality of the tested video corresponds to an achieved rate of 0.13 Mbps while the maximum is 4.72 Mbps. Additionally, between the minimum and the maximum rates, 18 segment rates are available for usage offering quality differentiation, thus ensuring the best fit at each segment served. In the UE’s side, we use VLC 3.0.3 Vetinari [24] for retrieving the DASH video streams. Similarly with the UDP scenario presented, the experiment’s duration is 90s. At \( t_0 = 0 \) we simultaneously initiate 7 VLC DASH streams from the UE4, which is connected to the gNB. We trigger the proposed framework every 30s and follow the exact same supplying capacity configurations as the UDP scenario. Finally, we alternate the adaptive buffer size in the DASH protocol and run the experimental scenario for two different values. The former adaptive buffer size is configured to 0s, which practically means that there will be no future segment buffering, while the latter configuration is set to 5s.

The collected results for each adaptive buffer configuration are depicted in Fig. 4b. There, we observe that the 7 VLC DASH streams initiated from the UEs, utilize almost all the capacity leased from the InP. Furthermore, we see that 3-5s are needed after the cell re-configurations for the MPEG-DASH to calibrate the best fitting rate. It is worth to be noted, that we don’t observe full cell’s utilization between 60-90s because even if all the 7 streams are served at the highest available rate of 4.72 Mbps, 33.08 Mbps are needed in total. Finally, it can be observed from Fig. 4b that the streams of 5s buffer size are served in a slightly lower quality. This is absolutely rational, as in the same experiment’s duration, 90 video segments are retrieved when we have 0s buffer size compared to 105 for 5s.

\(^3\)The interested user should refer to [12] for an analytical presentation of the theoretical game.
V. CONCLUSION

In this work, a novel framework for supporting dynamic charging rules was developed and presented. The proposed scheme extends the conventional PCF functioning and provides dynamic pricing when additional resources (slices) in case of capacity depletion should be leased by an InP. Through extensive experimentation, we assessed the performance of the proposed framework for both UDP and TCP additional traffic demands. We relied on the stable release of the FlexRAN implementation that currently supports the LTE-A 4G protocol stack to implement our work. Though it is being evolved to the next 5G protocol stack, the new releases are still under development. However, due to FlexRAN’s elasticity, the agents/master controller can be extended in order to support additional protocols. As a future extension, we examine the integration of WiFi, since the widely deployed IEEE 802.11 networks are anticipated to play an important role in the 5G network architectures.

VI. ACKNOWLEDGEMENTS

We acknowledge support of this work by the project “HELIx-National Infrastructures for Research” (MIS 5002781) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme “Competitiveness, Entrepreneurship and Innovation” (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). The research leading to these results has also received funding by the EU H2020 Programme for research, technological development and demonstration under Grant Agreement Numbers 857201 (5G-VEICTRI).

REFERENCES