Wireless Communications and Mobile Computing

Slicing in Modern Cellular Networks

Lead Guest Editor: Piotr Zwierzykowski
Guest Editors: Pei Xiao, Dejan Vukobratovic, and Anna Wielgoszewska

Slicing in Modern Cellular Networks

Wireless Communications and Mobile Computing

Slicing in Modern Cellular Networks

Lead Guest Editor: Piotr Zwierzykowski
Guest Editors: Pei Xiao, Dejan Vukobratovic,
and Anna Wielgoszewska

Editorial Board

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2019, Article ID 8685418, 2 pages
https://doi.org/10.1155/2019/8685418

Editorial
Slicing in Modern Cellular Networks

Piotr Zwierzykowski,[1] **Pei Xiao**,[2] **Dejan Vukobratovic**,[3] **and Anna Zakrzewska**[4]

[1]*Poznan University of Technology, Poland*
[2]*University of Surrey, UK*
[3]*University of Novi Sad, Serbia*
[4]*Nokia Bell Labs, Dublin, Ireland*

Correspondence should be addressed to Piotr Zwierzykowski; piotr.zwierzykowski@put.poznan.pl

Received 10 April 2019; Accepted 10 April 2019; Published 30 April 2019

One of the basic assumptions of the 5G network architecture is network slicing. It has the advantage of enabling, as a result of creation of logical networks (slices), simultaneous optimization of hardware and radio resources and provide network communication capability that is service-oriented. The ability to create virtual networks, dedicated to different users or services, allows for efficient dimensioning and management of allocated resources.

In this special issue of the "Slicing in Modern Cellular Networks", five high-quality papers are selected for publication whose authors consider typical challenges related to slicing in cellular networks, such as:

(i) resource management algorithms in logical 5G networks,

(ii) virtualization of 5G network components,

(iii) Cloud Radio Access Network (C-RAN),

(iv) hybrid architectures of 5G with other wireless technologies (e.g., SDN/NFV, satellite networks),

(v) analysis and modeling of 5G multiservice logical networks.

One of the articles is devoted to resource management algorithms in logical 5G networks. The article, entitled "Waveform Flexibility for Network Slicing," is written by Ł. Kułacz et al. The authors propose heuristic algorithms for waveform selection and frequency assignment in order to optimize the use of the spectrum and provider's infrastructure. The authors consider the possibility of cognitively adjusting the shape of the waveform to the requirements associated with various network slices jointly with the selection and allocation of the appropriate frequency bands to each slice. It is worthwhile to emphasize that in the proposed algorithms the assumed slice-specific Quality of Service (QoS) parameters are guaranteed. The proposal is verified and evaluated by simulation experiments conducted for four case studies covering several network deployment scenarios with a number of slices characterized by different QoS requirements.

Another article discusses virtualization of 5G network components. In the article "A Service-Oriented Approach for Radio Resource Management in Virtual RANs," B. Rouzbehani et al. propose a centralized cooperative mechanism of Radio Resource Management for Virtual Radio Access Networks based on aggregation and virtualization of all the available radio resources from different Radio Access Technologies (RATs). The presented virtual platform is responsible for service orchestration among Virtual Network Operators, enabling a definition of various services and policies, separately from vendors and underlying RATs. The virtual platform has the advantage of maximizing the virtual capacity utilization from different RATs in order to satisfy the specific QoS requirements for each service. The performance of the proposed model is discussed based on a simulation study. The results confirm that the model can capture the demanded capacity in keeping with the concept of proportional fairness.

L. M. P. Larsen et al. in one of the articles consider cloud RAN. The authors, in their article entitled "Fronthaul for Cloud-RAN Enabling Network Slicing in 5G Mobile Networks," present the Cloud Radio Access Network (C-RAN) architecture as a promising element which enables the introduction of virtualization and network slicing. In

particular, the authors pay special attention to the fronthaul network which links a distributed and a centralized unit of C-RAN. Proper deployment of a fronthaul network requires careful consideration of the trade-off between fronthaul bitrate, flexibility and complexity of the local equipment close to the user. In the numerical part of the article, the authors present the fronthaul range in relation to processing delay for exemplary LTE and Ethernet scenarios.

Another article is related to hybrid architectures of 5G with other wireless technologies. It is entitled as "A Biological Model for Resource Allocation and User Dynamics in Virtualized HetNet," is written by L. Ma et al.. The article investigates the dynamic-aware virtual radio resource allocation based on utility and fairness. The authors also propose a virtual radio resource management framework in which the radio resources of different physical networks are virtualized and form a common virtual resource pool. The virtual pool can be then used by mobile virtual network operators (MVNOs) to provide service to users. In the article, the authors propose a virtual radio resource allocation algorithm based on a biological model. In this model, the virtual resource allocation problem is formulated as a population competing problem, where users in the system are considered as the predators in nature environment, and virtual resources are prey to be hunted down by users. The algorithm is based on the Lotka-Volterra model, in which the authors introduced aggregated utility function and are considering together fairness and utility of the system. The authors also present the results of simulation experiments to verify that the proposed virtual resource allocation algorithm achieves a better trade-off between the total utility and fairness than the existing algorithm. It is worth emphasizing that the algorithm can also be utilized to analyze the population dynamics of system.

Finally an article proposes new solutions for analysis and modeling of 5G multiservice logical networks. M. Głąbowski et al. in their article entitled "Modelling of Multiservice Networks with Separated Resources and Overflow of Adaptive Traffic" propose a new method of determining traffic characteristics of multiservice overflow systems that carry adaptive traffic. The authors assume that adaptive traffic is subjected to the threshold compression mechanism in both primary and secondary resources that can be physical resources as well as virtual resources (e.g., slices). The model was elaborated on the basis of a generalization of Hayward's concept and its application to model systems with adaptive traffic with threshold compression. The presented method allows grade of service parameters (blocking probability, carried traffic, and network load) to be analytically determined and therefore can be used for optimal dimensioning of slices in modern mobile systems. The obtained calculation results are compared with the results of simulation experiments for a number of selected structures of overflow systems that service adaptive traffic, confirming the accuracy of the proposed analytical model.

We do hope that the selected articles will spark many discussions among the readers contributing to new research concepts and efficient implementation of slicing in cellular networks. We would like to thank everyone who contributed to this special issue, the authors for the interesting submissions, the reviewers for timely feedback, and the staff from the Editorial Office for their invaluable assistance.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Piotr Zwierzykowski
Pei Xiao
Dejan Vukobratovic
Anna Zakrzewska

Hindawi
Wireless Communications and Mobile Computing
Volume 2019, Article ID 6250804, 15 pages
https://doi.org/10.1155/2019/6250804

Research Article

Waveform Flexibility for Network Slicing

Łukasz Kułacz, Paweł Kryszkiewicz, and Adrian Kliks

Chair of Wireless Communications, Poznan University of Technology, 60-965 Poznań, Poland

Correspondence should be addressed to Adrian Kliks; adrian.kliks@put.poznan.pl

Received 29 November 2018; Revised 1 February 2019; Accepted 11 March 2019; Published 27 March 2019

Guest Editor: Anna Zakrzewska

We discuss the idea of waveform flexibility and resource allocation in future wireless networks as a promising tool for network slicing implementation down to the lowest layers of the OSI (Open Systems Interconnection) models. In particular, we consider the possibility of cognitively adjusting the shape of the waveform to the requirements associated with various network slices. Moreover, such an adjustment of waveform shape is realised jointly with the selection and allocation of the appropriate frequency bands to each slice. In our approach, the definition of the waveform, as well as the assignment of resources, is done based on the information about the surrounding environment and each slice requirement stored in a dedicated context-information database. In this paper, we present the key concept of waveform flexibility for network slicing, the proposed algorithm for waveform selection and resource allocation among slices, and the achieved simulation results.

1. Introduction

It has been almost twenty years since the concept of cognitive radio (CR) technology was introduced for the first time as a method of effective usage of spectrum resources [1]. It was assumed that a cognitive terminal or cognitive system should sense the spectrum (or better, sense the ambient environment) to locate itself in the surrounding transmission context. However, the sensing feature, which should allow for accurate and stable identification of unused frequency bands or ongoing transmissions, if applied by the stand-alone device, is said to be not accurate enough to guarantee the required level of quality [2–4]. Thus, in consequence, the researchers and engineers started thinking how to bypass this problem in practical applications, and the implementation of the dedicated context-information database appeared to be a viable solution [5]. Such a database is populated with, possibly, up-to-date and precise information about the ambient context. The cognitive terminal or cognitive system will make a decision about spectrum occupancy and the best way of using the spectrum resources based on its sensing capabilities supported by the information stored in the database. In this situation, it is worth mentioning two regulatory approaches that utilise dedicated databases, which may act as some sort of confirmation that this approach can be considered for practical systems. First, European Telecommunications Standards Institute (ETSI) has focused their efforts on the Licensed Shared Access (LSA) scheme, devoted mainly for 2.3 GHz band, where the LSA repository and LSA controller support the functioning of the cognitive radio based system [6]. Second, in USA, the Federal Communications Commission (FCC) considered a solution called the Citizen Broadband Radio Service (CBRS). In this case, the dedicated Spectrum Access System (SAS) uses databases for managing of priority access users [7]. In a nutshell, databases are widely treated as one of the technical enablers for practical deployments of highly flexible, cognitive radio based systems.

On the other hand, recent developments in the domain of network management and in the general functioning of communication networks have shown great benefits of wide virtualisation of network functions. The definition of virtual network functions, as well as a network orchestrator, allows for a precise split and separation between the underlying hardware and the functions that have to be delivered. The successful application of this scheme in wired networks has resulted in a great interest among researchers in adopting this technique also to the wireless network, mainly to the radio access domain. It is, however, not an easy task in general due to the great variability of the typical wireless channel. Nevertheless, the successful investigation of network

virtualisation is strongly connected with another concept, called network slicing. Based on various definitions of what a network slice is [8, 9], one may understand a network slice as an autonomous virtual network, being a part (slice) of a bigger communication network, which is created to fulfil specific needs (and requirements) and which can be managed separately and solely by the slice owner (tenant). In other words, specific network functions will be selected and delivered, so that the tenant is able to communicate autonomously within the created virtual network, being now aware of the underlying devices and technologies. When it comes to the Radio Access Network (RAN), the slicing concept is often referred to as RAN slicing [10], and quite often the presence of virtual radio resources is considered there [11]. Moreover, when discussing wireless slicing, one may refer to slice scheduling, Physical Resource Block (PRB) scheduling and traffic shaping [8] and, in general, inter-slice resource scheduling [12]. Interesting approach for inter-slice resource sharing based on graph colouring has been presented in [13], where the slices have been allocated in the Time Division Duplex (TDD) networks. The solutions proposed there assume that entire spectrum is allocated to the slice, and the traffic conditions are reflected in proper definition of uplink/downlink ratios. In [14] the authors proposed the application of so-called Network Slice Broker, the entity enabling mobile virtual operators to request resources in dynamic way via proper signalling.

Bearing the two above observations in mind, one may say that both cognitive radio and network/RAN slicing with network function virtualisation offer much flexibility to the network designers. However, as network slicing is more of an overall view of the communication network, cognitive radio is more focused on lower layers, mainly on the physical and medium access layers, and as such is more connected to RAN slicing. In fact, we claim that the functionality delivered by the cognitive radio technology can be treated as technical enabler for the practical deployment of the network/RAN slicing concept. In particular, we claim that the ability to dynamically adjust the waveform, as well as to adaptively assign the frequency band, is crucial for future implementation of the slicing concept. Following the idea of virtual resource blocks, one may imagine that the requirements for each slice may be specified in a generic way, meaning only the technology-independent constraints could be identified. For example, one may define a slice by defining the requirements on rate and latency, which may be translated into a specific set of virtual resource blocks, which in turn have to be mapped to specific (real) spectrum resources. In such a case, the cognitive radio technology can play its role by allowing a dynamic definition of the best waveform for data transmission, as well as by the proper association of the frequency band with the slice. In this paper we concentrate on the last functionality, and mainly we discuss the algorithms for adaptive selection of the waveform and the solutions for adaptive allocation of frequency blocks to the slices subject to fulfilment of their requirements on rate and Signal to Interference plus Noise Ratio (SINR). The algorithms are defined from the point of view of resource provider, i.e., the entity that delivers spectrum resources to various tenants who use slices for the delivery of their services. The spectrum provider wants to allocate appropriate frequency resources to specific slices in a way that maximises its revenue from spectrum licensing. It is also worth mentioning that the concept of waveform flexibility was first introduced in [15], where the primary system operating in the TV band was protected by the appropriate selection of the best waveform. This idea is extended in this work by adding the solutions for dynamic frequency allocation to the slice. Thus, the novelty of the paper is twofold; first, it extends the concept of waveform flexibility to network/RAN slicing, and, second, it proposes the algorithm for joint waveform and frequency band allocation among slices. The novelty of this manuscript can be summarised in the following way:

(i) First, we define the new concept of waveform adaptation and selection as part of the first, we define the new concept of waveform adaptation and selection as part of the network slice creation process, which is assumed to be fully dynamic and which maximises the resource provider revenue; we extend the work of [15], where the general idea of the waveform selection was initially discussed;

(ii) Second, we propose a joint scheme for waveform selection and resource allocation, and then we propose a two-step algorithm for network slice creation by spectrum and infrastructure provider;

(iii) Third, we have tested three algorithms for resource allocation that minimise the newly defined fragmentation coefficient.

The paper, being a sort of position paper, is organised as follows. First, we provide the scenario definition explaining the role of (a) spectrum and infrastructure provider (SIP) who offers its resources to (b) service provider (SP). Next, we present the network slice definition from the point of view of the lowest layers of the OSI stack model, which we applied in our research. In the following sections, we discuss the waveform selection idea and the frequency allocation problem, and we provide the proposals of three heuristic algorithms trying to solve these problems. The simulation results for four separate use cases are presented afterwards, and the entire work is summarised and concluded in the last section.

2. Problem Definition

In our work we consider the presence of a spectrum and infrastructure provider, whose assets include both infrastructure and spectrum resources with active licenses. The provider is interested in efficient usage of their available resources by dynamic creation of network slices for various types of end-to-end data transmissions. Moreover, their ultimate goal is to maximise the total revenue through maximising instantaneous spectrum usage and minimisation of various costs (in our case, it is the cost of consumed energy).

There is also a set of N prospective network operators (stakeholders, tenants) willing to deliver new services to end

users, but they do not want to invest in infrastructure and apply to the regulator for long-term licenses. Thus, these tenants would like to use the resources of the SIP by utilising the benefits originated by the network slice technology. These tenants are referred to as service providers to the end users.

In general, various slices may be created following the specific requirements defined by the interested stakeholders. One may refer to the recent work within 3GPP [16], where the network slicing is discussed as a vital part of the system architecture for the 5G systems. The functional description identifies the so-called Network Slice Selection Function (NSSF). The authors of [16] also inform that the network slice is defined within a Public Land Mobile Network (PLMN), and it should include the core network control plane and user plane network functions. Furthermore, within the serving PLMN, it should include at least either the new Next Generation Radio Access Network (NG-RAN) or the Non-3GPP Interworking Function (N3IWF) to the non-3GPP access network. It is claimed that network slices may differ for supported features and network functions optimisations and that the operator can deploy multiple network slice instances delivering exactly the same features but for different groups of end users. In that context, the so-called Network Slice Selection Assistance Information (NSSAI), which identified the network slice, consists of two fields: SST, standing for Slice/Service Type, defining the expected slice behaviour, and SD, Slice Differentiator, to distinguish among multiple network slices of the same SST. The standard [16] specifies three SSTs, mainly:

(i) for SST set to 1, the slice for 5G enhanced Mobile BroadBand (eMBB) type of traffic is considered;

(ii) for SST set to 2, the slice for handling Ultra Reliable Low Latency Communications (URLLC) is envisaged,

(iii) and finally, for SST set to 3, the slice for massive IOT (mIOT) type of traffic is proposed.

However, the standard does not specify the strict values of the features that have to be guaranteed by a specific slice, and this issue is still left open for implementation. The strict performance requirements for various scenarios (e.g., for high data rate and traffic density scenarios, for low latency and high reliability, and high accuracy positioning) are specified in [17].

Similar approach is presented by GSM Association in [18], where various exemplary slice requirements have been presented. One may identify the following suggestions (we present here only a subset of long list of various prospective applications):

(i) for augmented and virtual reality applications, the network slice shall consider various aspects of video codecs; for example, for strong interactive virtual reality schemes it is envisaged that the expected needed data rate could vary from 120 Mbps to 3.36 Gbps, the Round-Trip-Time (RTT) shall vary between 5 and 10 ms, and the acceptable packet loss is not greater than 10^{-6}. However, these values differ significantly if one looks at the requirements for weak-interactive virtual reality schemes, where the required data rate varies between 40 Mbps till 2.34 Gbps, and the allowable RTT is between 10 to 30 ms.

(ii) For automotive applications, there are not strict values provided, one may find the generic statement that mobile system shall provide ultra reliable and low latency communications between vehicles.

(iii) For the energy applications the requested bandwidth shall be around 1 kbps per user, the end-to-end latency shall be below 5 ms, the packet loss is less than 10^{-9}, and the availability is above "five nines".

(iv) For healthcare application, the definition of the important requirements is left for further investigation.

Thus, one may observe that there are no strict requirements associated with a generic slice of eMBB, mIOT, or URLLC traffic types in [18]. Similarly, in [19] the requirements put on various slices are defined in a highly generic way for the three fundamental traffic types, eMBB, URLC, and massive machine type communication, mMTC. Let us stress that the presented work concentrates on the ways for allocation of the physical resources certain slices (in particular, how the waveform can be selected and how the spectrum resources will be assigned), which is not a subject of standardization yet. One may observe that when the number of waveforms is fixed, the selection of the best one for given slice may be done in the analogous way as it is done nowadays for, e.g., selection of modulation and coding schemes.

As, theoretically, the number of various slices managed within one network is not limited, in order to make the analysis more tractable, we limit the number of different tenant types (which corresponds to different slice definitions) to three, mainly mIOT, eMBB, and URLLC types, as specified in [16]. The former one will be characterised by moderate latency and relatively low throughput, whereas the latter, by high requirements on the achieved rate. Let us stress, however, that the limitation on the number of slice types does not entail any requirements on the number of tenants of each type. It means that the SIP may create a few network slices of both types.

In the context of OSI layers, there is still debate as to how the network slices should be incorporated into the lowest layers (mainly physical and medium access control). From this point of view, we assume that the slice itself will correspond to the set of requirements that have to be guaranteed by the created virtual network, and the fulfilment of these requirements may be achieved by adaptive definition of various parameters. In particular, we consider that the SIP (i.e., the entity responsible in our case for network slice creation) will be able to adapt the waveform shape to the slice definition and to instantaneous channel conditions. Moreover, a dedicated algorithm for spectrum band assignment to the slice will be run on the network part in order to maximise the revenue of the spectrum provider under the constraint of permanent fulfilment of slice-specific requirements. In our analysis, we assume that

the network slice will be created at least in a minute scale; i.e., the assignment of frequency bands to the slices is not a problem of classical resource scheduling applied to cellular networks (like proportional fair or round-robin), but more of creating an appropriate frequency allocation plan among base stations. Moreover, once the frequency band is associated with a certain slice, the dedicated scheduler may run within that slice to allocate resource blocks among the slice users. It also means that the SIP will try *to pack* various tenants in such a way that the spectrum utilisation is maximised, and such spectrum assignment is valid for a relatively longer time.

3. Slice Definition

As mentioned above, without the loss of generality, we have limited our analysis to three most common types of slices: a low rate slice for mIOT type of traffic, a high rate one for eMBB service providers, and one called URLLC for low latency schemes. From the perspective of lower layers of the OSI model, the slice will be characterised by a set of parameters describing the requirements that must be guaranteed by the SIP to the tenants. Following the discussion in previous section, as well as the observation presented in [16–19], we agree that there is not one strict definition of the set of requirements for a certain network slice. Assuming the application of network virtualisation and cognitive ratio terminals, it will be the role of the SIP to map the requirements onto the available physical resources (e.g., spectrum bands) so that the assumed quality of service requirements within the network slice is satisfied. Having this in mind, the mIOT slice may be in general defined as a network where the rate of some tens or hundreds of kilobits per second per one user is acceptable, and there are no strict requirements on data delay; i.e., delays typical for cellular networks are acceptable. It is also expected that simple and mature technologies may only be applied, as the cost per one chip, as well as the overall energy consumption, should be minimised. The possible physical layer technologies fulfilling these requirements are, for example, the traditional GSM transmission scheme, or such solutions as Long RAnge (known as LoRA), SigFox, Long-Term Evolution for Machines (LTE-M), or Narrowband Internet of Things (NB-IOT) [20]. For our further analysis we assume that the Gaussian Minimum Shift Keying (GMSK) transmission occupying 200 kHz will be applied. We also assume that the number of active channels may vary in time.

Analogously, the eMBB network will require the rate of a few megabytes or more per one user, and there are again no rigorous requirements on the network slice latency. The end-user devices are expected to have relatively high processing power, so more advanced technologies such as Long-Term Evolution (LTE), LTE-Advanced (LTE-A), or New Radio (NR) may be considered. However, the energy consumption should always be kept at the lowest possible level. Similarly, the URLLC network may be characterised by extremely strict requirements on latency and reliability and relatively high requirements for the traffic rate.

4. Waveform Selection Idea and Frequency Allocation Problem

As stated above, the key research problem is to find the way for spectrum usage maximisation (and, in turn, revenue maximisation due to spectrum leasing) by, first, adaptive selection of the waveform shape within the slice and, second, through smart allocation of frequency band to each slice. It can be easily observed that the selection of the waveform immediately entails the occupancy of certain spectrum by the slice, and thus it has a potential impact on the frequency allocation algorithm. At the same time, each waveform can be characterised by its computational complexity (or more precisely, the number of required mathematical operations per one transmitted bit), as defined in [21]. Thus, more advanced waveforms (such as noncontiguous, filtered multicarrier schemes) may guarantee narrower spectrum at the expense of advanced and computationally intensive processing. Narrower spectrum associated with given waveform (thus, with slice as well) may cause more slices of the same type to be created within a certain frequency band W. But on the other hand, a service provider (i.e., the entity for whom the slice is being created) may be interested in as simplified processing as possibl and will accept advanced waveforms only in specific situations or at a higher price. Moreover, more advanced waveform needs typically more sophisticated processing and power consumption; thus it may have slight but negative impact on the overall end-to-end delay of the link. In consequence, assuming that there are N various waveforms to select from, one may calculate the corresponding transmittance of each waveform $H_n(f)$ for n=1,2...N, and one may compute the number of necessary mathematical operations (see [21]), denoted as β_n, and the associated processing power $P_n^{(\beta)}$. In such a case, one may think about the joint optimisation algorithm that would find the best solution; i.e., it will maximise the SIP's revenue ψ by the proper selection of the waveform and frequency band. Various factors can be considered in the definition of the revenue, such as requested bandwidth $B_s \leq W$, network slice lifetime T_s (i.e., the requested time for which the slice has to be maintained), requested coverage area A_s, and expected end-to-end delay θ_s. In that context, the highly generic formula for the revenue could be defined as function of the abovementioned factors, i.e., $\psi(B_s, T_s, A_s, \theta_s, P_n^{(\beta)})$. Clearly, the more resources (e.g., bandwidth, allocated time, covered area) are requested, the greater revenue shall be expected. At the same time, the more advanced waveform is selected, the more processing power and costs at the SIP side is envisaged. Depending on the exact definition of the revenue function, appropriate waveform selection and resource allocation algorithm will be selected. In our approach, we assume the following:

(i) first, there are no priorities between service providers,

(ii) second, the SIP acts in a greedy fashion; i.e., there is a willingness to serve each incoming request from service provider, and only the lack of resources prevents the SIP from creation of a new slice

(iii) third, the considered revenue of the SIP increases with the requested rate, and decreases with the processing cost

(iv) finally, the SIP processes the incoming requests immediately at the time they appear in the system, and there is no queue for nonserved SPs.

In consequence, the SIP is interested in delivering services at the lowest cost (based on the third assumption), so it will try to select the simplest, yet spectrally efficient, waveforms minimising the processing costs. Please note, also, that the SP is also interested in utilisation of the simplest waveforms, as the selection of advanced solutions by the SIP entails more severe processing also at the SP devices. Thus, we have decided to split this problem into two subproblems, knowing that the optimal solutions cannot be achieved, but such an approach makes the problem practically tractable. In the first step, we check the possible set of waveforms that could theoretically be assigned to the new slice, and from this set we select the one that guarantees the simplest processing (i.e., it achieves the lowest energy consumption due to the lowest number of required operations per one transmitted bit). Once the waveform is selected, we try to find the best allocation of the considered slice in the frequency domain. We assume that the entire considered frequency band of bandwidth W is split into C equal channels, each of bandwidth W/C. Adjacent unused channels (vacant channels) create channel blocks, which are interwoven with the channels already allocated for network slicing. If there are $C(i)$ adjacent vacant channels in block i, the total bandwidth of the channel block i is equal to $C(i)W/C$. For example, let us assume that the band of 20 MHz is split into 40 channels each of 500 kHz bandwidth. Originally, when no slice is created and all channels are unused, there is one block of 40 channels. When two channels in the middle (e.g., the 19$^{\text{th}}$ and 20$^{\text{th}}$ channel) are assigned to a particular network slice, there are two channel blocks that can be further allocated to new network slices. The first block consists of 18 channels, and the second block consists of 20 channels. These two blocks in the frequency domain are interwoven with a 2-channel-wide spectrum block.

In the following, we describe these two phases in more detail. Let us also mention that in our analysis we initially assume that the service providers (tenants) are interested in delivering their services over possibly wide areas with the highest possible signal quality. Thus, we assume a fixed and constant value of the transmit power in each slice. However, it is worth mentioning that the potential introduction of power adaptation creates further degrees of freedom in system design but is intentionally left for further study.

4.1. Step I: Waveform Adaptation. Following the assumption mentioned in the previous sections, the goal of this phase is to find the simplest waveform, which will allow for achievement of the requirements defined in the request delivered by the SP. It can be mathematically represented as follows:

$$\min \quad \sum_{n=1}^{N} \varepsilon_n P_n^{(\beta)}, \tag{1a}$$

$$\text{s.t.} \quad R \geq R_{\text{req}}, \tag{1b}$$

$$T \geq T_{\text{req}}, \tag{1c}$$

$$A \geq A_{\text{req}}, \tag{1d}$$

$$\varepsilon_n \in \{0, 1\}, \tag{1e}$$

$$\sum_{n=1}^{N} \varepsilon_n = 1, \tag{1f}$$

$$\bigwedge_{m \in M} \sum_{n=1}^{N} \varepsilon_n P_{l,m}^{(n)} \leq \check{P}_{I,m}, \tag{1g}$$

where R_{req}, T_{req}, and A_{req} represent the requirements on rate, time, and area specified by the SP in its request for network slice creation and R, T, and A are the assigned resource corresponding to rate, time, and area. ε_n is a Boolean variable for nth waveform. Moreover, assuming that there are already M created network slices, $P_{l,m}^{(n)}$ defines the total interference power induced into the other already served system m when the n-th waveform is selected, and the value of the interference power cannot exceed the maximum acceptable interference for that system denoted as $\check{P}_{I,i}$. This aspect is discussed in detail in the following sections.

The initial concept of dynamic waveform definition was originally discussed in [15]. It was observed there that contemporary wireless communication systems utilise various adaptation procedures in order to improve the performance of wireless data delivery. Various transmit parameters may be adjusted in the wireless communications systems today, such as the selection of various modulation and coding schemes (MCS) or adjustment of transmit power via open or closed power control loops. In [15] we proposed to consider waveform flexibility, where the cognitive terminal may decide to select one of four available waveforms for data transmission, mainly, traditional Orthogonal Frequency Division Multiplexing (OFDM) signalling or Filter Bank-Based Multicarrier (FBMC) schemes [22] and their non-contiguous versions, NC-OFDM, and NC-FBMC [23]. The ultimate goal in that work was to use vacant spectrum while protecting the primary transmission, mainly the Digital Video Broadcasting–Terrestrial (DVB-T) signal.

In this work, we focus on horizontal sharing scheme, where the set of frequency resources (licensed to one party, a SIP in our case) will be shared among all interested stakeholders (the SPs) and no priorities are considered between them. The dynamic mechanisms are considered here as support for slice creation at the physical and medium access layers, and these slices are assumed to be of identical importance. It means that all existing transmissions have to be protected, and the new network slice will be created only in a case, when no harmful interference will be induced into any of the already existing slices. Moreover, limited set of available waveform are taken into account, and mainly we have selected the Gaussian Minimum Shift Keying (GMSK) scheme as suitable for the mIOT slice and contiguous and noncontiguous multicarrier signals (OFDM and NC-OFDM) for the eMBB one. The noncontiguous multicarrier scheme

has the advantage that it can efficiently utilise even highly fragmented spectrum, which is interweaved with other transmissions. The main cost of such approach is the increased number of operations to be performed per one second and increased control information overhead, when compared to the classical OFDM scheme. Let us remember that the ultimate goal of the spectrum and infrastructure providers is to maximise their revenue, while minimising any unnecessary costs. Thus, we claim that depending on the situation (i.e., how much interference can be induced to neighbouring systems) the transmitter should select the simplest possible waveform shape that guarantees the achievement of the defined requirements. Thus, it should first prefer to select the contiguous transmission scheme (i.e., OFDM), and if the selection of the contiguous band is not be possible, the noncontiguous versions of this multicarrier scheme can be applied. Such an approach is justified, as the overall complexity of the OFDM case is much lower than that of NC-OFDM.

Assuming that the tenant is capable of using all waveforms, the following algorithm for waveform selection phase will be applied:

(i) for given requirements (e.g., on average rate), check the possibilities of application of the simplest waveform (in our case GMSK); if it is possible, go to phase 2 of the algorithm and try to assign frequency resources for GMSK-based slice; if this is possible, the algorithm is finished and slice is created;

(ii) if the achievement of assumed technical requirements is not possible with GMSK or it is not possible to assign frequency resources, check if it is possible to apply OFDM signalling and, if yes, try to assign frequency bands for OFDM slice;

(iii) when the achievement of assumed requirements is still not possible, follow the same procedure with NC-OFDM scheme.

4.1.1. Interference Analysis. The assignment of specific frequency bands to the slice is strongly connected with the need to keep the mutual (also aggregated) interference below acceptable level. In our scenario we assume the same hierarchy of all slices, thus the problem of mutual interference has to be considered within each pair of coexisting slices (systems). Let us recap that the interference observed between neighbouring frequency channels is caused by the imperfections of the real transmitter (such as the characteristics of the impulse shaping filter and nonlinearities at the radio front end leading to Out-Of-Band Emission, OOBE) and receiver (e.g., non-ideal selectivity of the reception filters). The problem at the transmitter side is addressed by providing (typically provided precisely in the standards) a definition of the spectrum emission mask (SEM), which specifies the requirements on the minimum transmitted signal attenuation at a given frequency. At the same time, the design of the reception filters at the receiver side (thus, its transmittance) is typically left to manufacturers. Although the minimum selectivity level can be specified in standards as well, its characteristics (averaged over various designs) can be retrieved by measurements. These two sources of imperfections illustrate two sources of interference observed between adjacent (in the frequency domain) wireless systems. In particular, SEM specifies how much power may be introduced to the neighbouring channels if the transmitter sends a signal with a given power. Analogously, the frequency response of the effective reception filters defines the amount of unwanted power intercepted by the receiver from the neighbouring bands. In the literature, these two phenomena are described mathematically by

(i) Adjacent Channel Leakage Ratio, ACLR, which describes the ratio between the power transmitted in the nominal band of the system and the power observed in the adjacent band

(ii) Adjacent Channel Selectivity, ACS, which informs about the ratio of receiver filter attenuation on the band of interest and filter attenuation on the adjacent channel frequency.

Typically these two factors are often represented jointly as the Adjacent Channel Interference Ratio, ACIR, defined in linear scale as

$$\frac{1}{ACIR} = \frac{1}{ACLR} + \frac{1}{ACS}. \tag{2}$$

To be more specific, the total interference power introduced to a given system can be calculated in the frequency domain as

$$P_I = \int_{-\infty}^{\infty} PSD(f) \left|H^{TR-RX}(f)\right|^2 \left|G^{RX}(f)\right|^2 df, \tag{3}$$

where $PSD(f)$ is the power spectral density at the output of the interfering transmitter antenna (including the power allocated to given symbols/subcarriers, symbol shaping in digital domain, digital-analogue converter characteristic, and frequency response of the transmit—TX—antenna), $H^{TR-RX}(f)$ is the channel frequency response (considering channel attenuation and multipath propagation effects), and $G^{RX}(f)$ is the frequency response of an effective receiver filter (including the reception—RX—antenna characteristics, analogue frontend characteristics, and demodulator characteristics). The formula (3) provides precise method of interference calculation. However, in many cases it is not practical as the current channel fading is unknown. From this perspective it is not needed to limit instantaneous interference power, but it is enough to constrain expected interference power, i.e., E[PI] assuming the only random variable is channel coefficient under fast fading. Therefore,

$$\begin{aligned} E[P_I] &= \int_{-\infty}^{\infty} PSD(f)\, E\left[\left|H^{TR-RX}(f)\right|^2\right] \left|G^{RX}(f)\right|^2 df \\ &= \alpha \int_{-\infty}^{\infty} PSD(f) \left|G^{RX}(f)\right|^2 df \end{aligned} \tag{4}$$

assuming that mean channel coefficient power $E[|H^{TR-RX}(f)|^2] = \alpha$ is constant over some period and

independent of f. This is true, e.g., for Rayleigh fading. It can be argued that for sufficiently rich multipath propagation environment and sufficiently wide considered bandwidth the channel attenuation at some frequency will be connected with channel amplification at some other frequency making the expectation close to real instantaneous interference power. We do not claim that the wireless channel is flat although locally, at a given frequency f, it can be treated as flat (e.g., this is a reason for vast usage of OFDM modulation nowadays). As such, the wireless channel can be characterised by a single coefficient α being reversely proportional to the channel path loss giving

$$P_I = \alpha \int_{-\infty}^{\infty} PSD(f) \left| G^{RX}(f) \right|^2 df. \quad (5)$$

Defining the total power transmitted from the interference source as

$$P_{TX} = \int_{-\infty}^{\infty} PSD(f)\, df \quad (6)$$

ACIR, being the ratio of interference power at the antenna of the victim receiver to the effective interference deteriorating its performance, is defined as

$$ACIR = \frac{\alpha P_{TX}}{P_I}. \quad (7)$$

This definition is very useful in calculating the effective interference power between slices as

$$P_I = \frac{\alpha P_{TX}}{ACIR}, \quad (8)$$

where ACIR measures the interference coupling between the transmitter and receiver, independent of the transmission conditions, and transmission power. The assumption about ACIR independence from TX power is typically used, although it can be incorrect if nonlinear effects take place in the transmitter or receiver; e.g., frontend saturation in the interfered device is possible if the interfering system is located in a relatively near distance and is transmitted with relatively high power. There are various ways of obtaining transmitter/receiver frequency characteristics:

(i) Based on measurements: In this case all characteristics of transmitter/receiver are assessed during all processing phases (both analogue and digital). However, the obtained characteristics are specific for a given transmitter/receiver pair and the signal configurations used during measurements.

(ii) Based on standards: In this case, the worst case interference generation/rejection is specified. Although any produced device has to obey the specified limits, it will result in the lowest possible ACIR between systems.

(iii) Based on transmitter/receiver modelling: It requires knowledge about the modulation/demodulation used. Although ACIR in many scenarios can be calculated accurately, it requires some assumptions about the analogue frontend.

In our analysis, we consider the presence of two classes of slices. The first one utilises GMSK modulation (it is the class consisting the mIOT traffic; thus mIOT slice is part of it), whereas the second allows for the adaptive adjustment of the multicarrier signal by using either OFDM or its noncontiguous version (it is the class tailored for eMBB and URLLC traffic). The considered slicing model assumes all OFDM-based slices are orthogonal to each other (in order to reduce mutual interference to zero). It can be simply obtained if the transmitter spans the whole available bandwidth providing independent subcarriers to each OFDM-based slice. However, the mutual interference between GMSK and OFDM has to be considered. The GSM transmitter and receiver characteristic are based on standard-based receiver rejection characteristics in [24] and standard-based transmitter characteristics [25]. The calculation method is based on the method presented in [23]. The GSM-GSM ACIR calculation is based on the abovementioned transmitter/receiver characteristics with the calculation carried according to (5).

4.2. Step 2: Frequency Allocation. In the first step of the procedure, we select the best waveform, taking into account the requirements defined in the slice, in particular, the limitations regarding the available waveform shape and the constraints on the consumed energy. Once the prospective waveform is defined, the crucial point is to select the appropriate frequency band, such that the instantaneous spectrum fragmentation is minimised. We claim that it is easier to allocate a new network slice in the physical layer (mainly by assigning spectrum resources to the spectrum slice) if the spectrum is not fragmented; i.e., the vacant channels are wide and are not interwoven with ongoing transmissions. In that sense, the worst case is when the set of occupied channels and vacant channels creates a comb-like form in frequency domain. In such a case, only very narrowband channels are available; thus either slice that requires narrow bands can be allocated, or sophisticated noncontiguous multicarrier schemes have to be applied (i.e., where many narrow gaps in the spectrum are used for data transmission).

In order to provide the maximum flexibility to the SIP, it should have the ability to allocate spectrum to various slices with possibly minimum computational load. One may observe that the more fragmented spectrums (i.e., the more narrowband gaps of unused spectrum in the considered band), the less degrees of freedom in allocation of resource do the new slice. Spectrum fragmentation may result in a situation, when some SP (which are not capable in usage of advanced waveforms, such as noncontiguous schemes) will be not served by SIP, although theoretically SIP will have enough resources for such a new slice. Thus, the frequency allocation process should prefer such allocation schemes that will minimise the spectrum fragmentation or, in our approach, maximise the fragmentation coefficients. In our experiment, we have defined the fragmentation coefficient μ as follows:

$$\mu = \frac{\sum_{i=1}^{S} C(i)^2}{C^2}, \quad (9)$$

where S is the number of unused spectrum blocks, $C(i)$ defines the number of channels within the i-th block (i=1,. . ., S) of vacant channels, and C represents the total number of channels in the considered spectrum range W. This coefficient equals 0 when the whole spectrum is occupied and 1, when no channel is used. In order to increase the importance of wider bands, the number of channels within each block is squared. For example, if there are two 3 MHz wide blocks of unused channels, it is a worse case when compared to the situation when the blocks are of the width of 1 MHz and 5 MHz, respectively. Please note that, without the power coefficient set to 2, these two exemplary cases will have the same value of the μ coefficient. Furthermore, please observe that the assumption that the SIP wants to maximise the spectral efficiency guarantees that the algorithm will not try to keep the fragmentation coefficient close to zero by not allocating the spectrum to anyone. In other words, each new request for slice creation has to be realised only if there is a space in the frequency domain.

Having in mind the definition of the fragmentation coefficient and knowing that the n-th waveform is chosen, the optimisation goal can be defined as

$$\max \quad \mu, \tag{10a}$$

$$\text{s.t.} \quad R \geq R_{\text{req}}, \tag{10b}$$

$$T \geq T_{\text{req}}, \tag{10c}$$

$$A \geq A_{\text{req}}, \tag{10d}$$

$$\bigwedge_{m \in M} P_{l,m}^{(n)} \leq \check{P}_{I,m}. \tag{10e}$$

We have proposed and tested three separate pragmatic algorithms for frequency assignment to network slices, and the three are discussed in detail below:

(i) *Method A* (also marked later as *linear*)*:* in this approach, the new slice is allocated at the first possible position in the spectrum starting from the lower bound of the available frequency band; in other words, once the waveform is selected, the algorithm searches for the first available position, taking into account the requirements on ACIR and SIR. To put it simply, the algorithm will assign the first available channel (or channels) from the left side of the available spectrum band, for which the requirements are fulfilled.

(ii) *Method B* (also marked later as *sinr*) also allocates the spectrum starting from the lower bound of the available frequency range, but it modifies method A in such a way that it maximises the value of final SIR in the system (i.e., observed after frequency allocation). As SIR will increase as the distance between the signal spectra in the frequency domain also increases, the following behaviour of the algorithm is expected. In particular, in the first step, the algorithm will allocate to the slice the frequency band on the first available channels and, in the second step, the new network slice will be allocated mostly as far as possible (in spectrum domain) from the existing transmission (therefore, on the right bound of the available frequency band). If the third slice is created, it will be allocated in the middle of two existing transmissions in the frequency domain.

(iii) *Method C* (also marked later as *fragmn*): as the two previous methods allocate the spectrum based on SIR criteria, the third method follows the brute-force search strategy with regard to the fragmentation coefficient μ. In other words, it searches for the best allocation of frequency resources which maximise the fragmentation coefficient μ.

Please note that in general SINR value should be considered here. However, one may observe that the proposed methods will be applied within one base station, and in consequence the impact of noise and path loss observed in each location of the covered area will be the practically the same for each slice (i.e., for given location the path loss and noise will be the same for each slice, if these slices are created within the same base station). Knowing that 1/SINR=1/SNR+1/SIR, we may remove the SNR component in the frequency allocation step, as it will not impact the result. Of course, in the evaluation of the interference between the cells and, for rate calculations, the presence of thermal noise is considered.

5. Role of Context-Information Databases

One of the key concepts that gains momentum in recent years in the domain of wireless networking is the efficient provision of access to dedicated databases that store various pieces of information about the environment [15]. We call them context-information databases, CI DB, and, in our case, such a database will be used by the central coordination entity used for network slicing management. The database may be populated with the following data defining the transmission context:

(i) a coverage map of the existing transmission (associated to the slices) that has to be protected from interference. The new slice has to be created in such a way that the existing transmissions are protected.

(ii) the specificity of each slice, such as the definitions of spectrum emission masks, possibly, approximations of the reception filter characteristics, historical values of computed ACIRs, and SIR.

(iii) past decisions to certain SP requests.

Having access to such information, the SIP may select the best transmit opportunity for each new slice. However, the way that the database is populated and the way the information should be stored is left for further study.

6. Simulation Results

In order to verify the performance of each proposed method, we have carried out extensive computer simulations in two

Table 1: Outage probabilities achieved for three proposed methods: use case 1.

Method	Outage probability			
	$(\xi, \omega) = (30,70)$	$(\xi, \omega) = (20,50)$	$(\xi, \omega) = (20,20)$	$(\xi, \omega) = (20,100)$
A	0.0950	0.1818	0.09504	0.0950
B	0.1178	0.1985	0.0214	0.2982
C	0.0840	0.1546	0.0840	0.0840

use cases, which are described below. The performance of the considered methods is measured twofold:

(i) First, by computing the fragmentation coefficient, i.e., after each frequency allocation we measure the value of the fragmentation coefficient and present it in the form of a Cumulative Density Function (CDF), which is equal to the probability that fragmentation coefficient μ is below some certain threshold μ_0, i.e., $\Pr(\mu < \mu_0)$. In the first use case, the fragmentation coefficient is measured per one base station, and, in the second use case, it is presented in various contexts (e.g., per entire network and per base station).

(ii) Second, by computing the outage probability, i.e., the probability of a situation that a new request for network slice creation is not realised due to the lack of resources.

6.1. Use Case 1: Single Base Station. In the first use case, we focus on network slicing within one separate base station; i.e., we assume that the SIP is in possession of one mast and offers the creation of various network slices using its resources. This particular example can be also understood as a case where each base station in a network is treated as an autonomous entity.

The simulation parameters have been set up as follows:

(i) The bandwidth of the considered spectrum for network slice allocation is set to $W = 10$ MHz, and the central frequency was set to 3.5 GHz (thus, the frequency range considered for simulation is from 3.45 GHz up to 3.55 GHz)

(ii) The band W is split into $C = 20$ channels, each of 500 kHz of bandwidth

(iii) The new process for appearance of a new request for network slice creation is modelled as an exponential one with the intensity of ξ assumed time units (i.e., the mean value was set to ξ time units); the definition of the slice duration (i.e., the time for which the slice is needed) is also realised as an exponential random variable with the mean value of ω time units; in this experiment we have verified the following cases: $(\xi, \omega) = \{(20,20), (20,50), (30,70)\}$,

(iv) The results have been collected over 100000 separate scheduler decisions runs, and the ultimate minimum requirement to stop the stimulation was the achievement of 1600 positive slice allocations in the whole simulation run

(v) The appearance of the requests for two considered network slices is equally probable.

The achieved results for three discussed methods are shown in Figure 1, where we compared the CDF of fragmentation coefficient μ for various setups of (ξ, ω). As the fragmentation coefficient should be maximised (i.e., we want to avoid the situation where the spectrum is highly fragmented), the performance of the tested method is better when, in simple terms, the CDF is shifted to the right in the plot.

One may observe that, for low and high values of μ_0, which corresponds to situations where the frequency spectrum is almost empty or fully loaded, respectively, all methods will achieve similar performance. The greatest performance improvement is observed in midrange of values of μ_0. The achieved results show that method C achieves usually the best results in the midrange of μ_0 regardless of the slice setup (ξ, ω). However, the most important conclusions from the achieved results are that the performance of the method significantly depends on the considered use cases. When the slice durations is relatively low (see Figure 1(c)) or high (see Figure 1(d)), the observed gain is rather negligible. Let us note that these results are achieved for the case when the noncontiguous waveform is not allowed for selection. As it will be observed in the following sections, the inclusion of these advanced schemes can lead to significant gains. Moreover, the CDF of fragmentation coefficient has to be analysed jointly with other metrics, such as outage probability—the achieved results are shown in Table 1. Outage probability reflects here the situation that there will be no resource allocated for a given slice. The value of this probability is calculated as the number of situations, where there was no resource allocated to a given slice over all simulation runs. One may observe that in three cases (i.e., for long slice duration) the worst performance is observed for method B, whereas the lowest outage probability is achieved mostly for method C. In other words, the outage probability will be minimised in such a scenario when the algorithm tends to minimise the spectrum fragmentation. Interestingly, in case when the slice duration is low, the best results are achieved for method B.

Let us notice that both results (CDF and outage probability) show the performance of the entire solution, i.e., waveform selection and resource allocation steps. However, our observations show that in the considered scheme the algorithm in most cases selects mainly OFDM as the optimal waveform, and the noncontiguous version of it was chosen rarely; thus the above figure as well as the table below rather shows the comparison of the three frequency allocation methods. In order to observe the impact of the waveform selection phase, we have analysed new use case no 3 (see

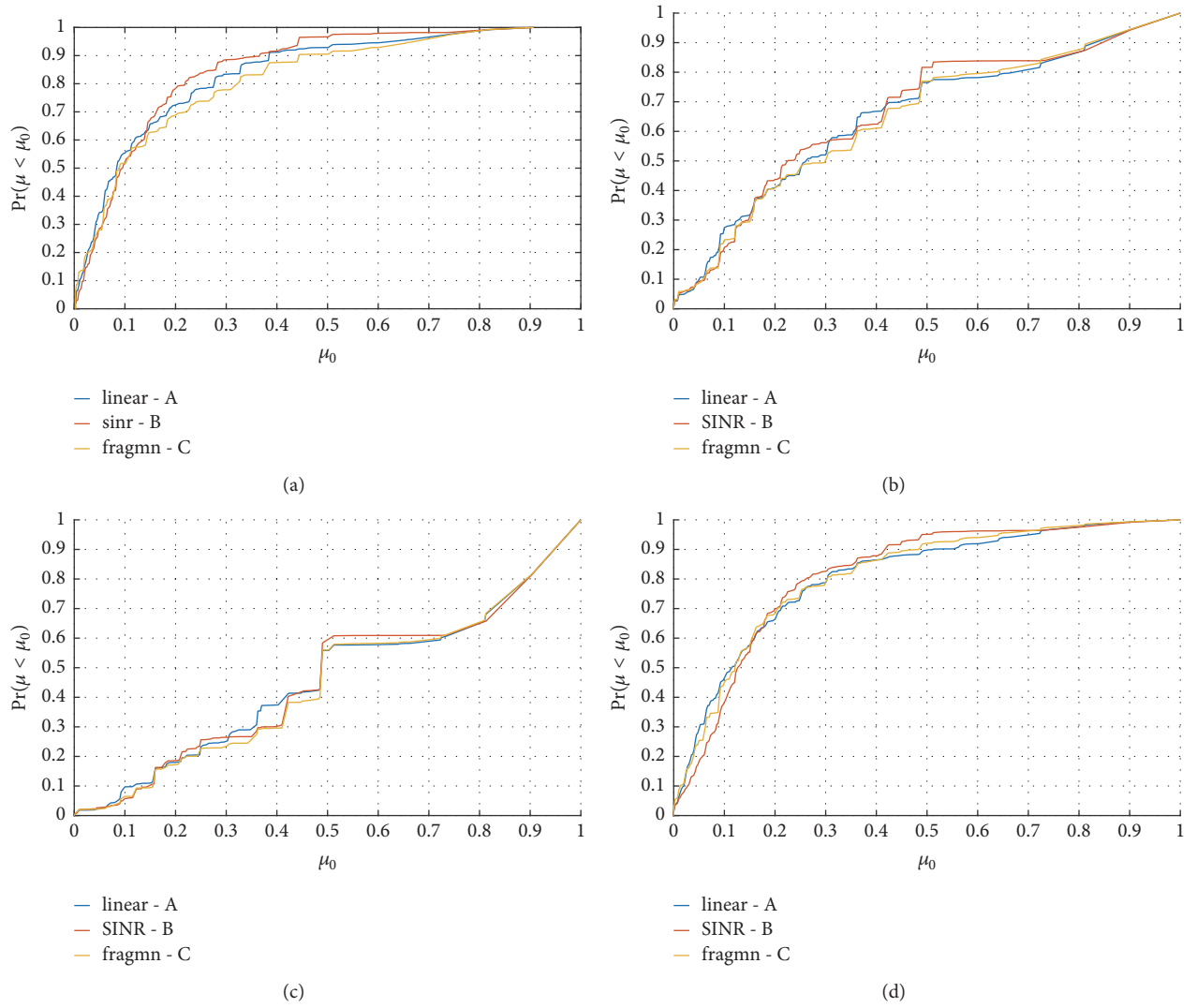


Figure 1: CDF of fragmentation coefficient observed at single station (both eMBB and mIOT slices can be deployed) for various setups of (ξ, ω): (a) (20, 50), (b) (30, 70), (c) (20, 20), and (d) (20, 100).

below), where the role of this phase of the proposed solution is dominant (please see the following section).

6.2. Use Case 2: Fragment of the Network. In the second use case, we are considering three base stations being part of a bigger network, as shown in Figure 2. We simulate the cells in the form of hexagonal regions and we assume that the potential interference originated from the outer cells (e.g., from the so-called first circle around this network fragment) is low enough to be excluded from computations. Such a simplification is tractable, as the presence of similar, low-power interference originated from the outer cell will have identical impact on all methods and does not influence the relations and correlations between the methods.

In order to adjust the three proposed methods to the new scenario, we have applied the following modifications when compared to the previous use case. Mainly, in order to make a frequency allocation decision, not only do SIR between adjacent systems in the frequency domain have to be considered (as it was done in the first use case) but also the requirements on resultant Signal to Interference plus Noise Ratio (SINR) values observed at the edges of the cell. Thus, the key extension of methods A, B, and C described above is that while allocating a certain frequency to the slice, the algorithm not only checks the SIR requirements but also computes the potential impact of the new frequency assignment on the SINR value observed at the cell edges. In practice, the SINR requirement has to be fulfilled at each point of the cell edge; in our simulation, however, we consider only selected points for evaluation, which are marked by red dots in Figure 2.

The simulation setup was in general the same as in the first use case, but the following additions have been included:

(i) as the SINR metric has to be computed, we assumed the propagation model defined in the WINNER project, which is designed for micro cell deployment in an urban scenario (so-called C2 option);

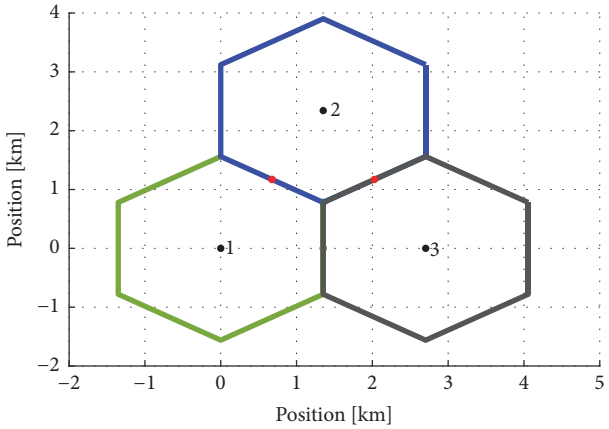


FIGURE 2: Topology of the considered fragment of the network.

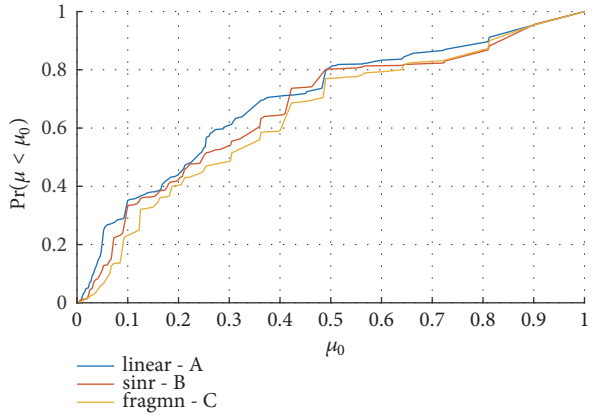


FIGURE 3: CDF of fragmentation coefficient observed at the first base station (both eMBB and mIOT slices can be deployed).

(ii) furthermore, the base station and end-user terminal are assumed to be deployed at the heights of 30 m and 2 m, respectively;

(iii) we assume that in both slices the transmit power is fixed and the power spectral density is set to 0 dBm per 5 MHz;

(iv) noise power was calculated for the temperature of 20 degrees Celsius;

(v) the cell edge contour is defined by the line where the observed SNR is equal to 5 dB over thermal noise; thus in the given scenario, the cell radius was equal to approximately 1.56 km;

(vi) the required SIR level between any two systems in the frequency domain was set to 26 dB;

(vii) and, finally, the required observable worst case SINR level, measured in the four positions marked in Figure 2 by black dots, is to be above -1dB.

Moreover, we have assumed that all base stations are technically capable to deploy slices of eMBB traffic type, but only two of them are able to generate signal suitable for the mIOT slice. In other words, a mIOT network slice may be created only over two base stations, whereas an eMBB network slice may be created over the entire network. Such a scheme is selected to illustrate the flexibility of slice creation; i.e., the tenant (SP) may be interested in delivering its services only in some specific area, and the SIP has to manage all requests for slice creation coming from all tenants. In consequence, when a new request for network slice creation appears, the algorithm tries to create the network slice over the entire area that supports slices of a given type. Such a need generates some new design issues; i.e., due to the mutual interference between the adjacent cells, the same frequency resources cannot be typically applied over the network area. The entire problem of frequency allocation now becomes highly similar to various solutions from the domain of frequency planning.

The achieved results ale plotted in Figures 3, 4, and 5, where the CDFs of fragmentation coefficients are shown per each base station and in Figure 6, where the CDF is calculated for the entire network. By analysing these figures one may

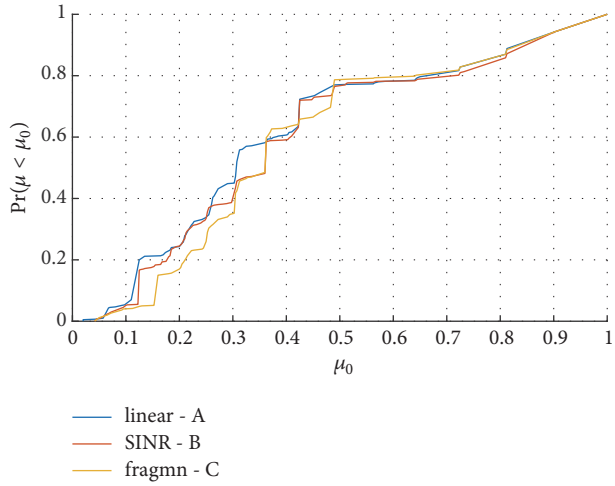


FIGURE 4: CDF of fragmentation coefficient observed at the second base station (both eMBB and mIOT slices can be deployed).

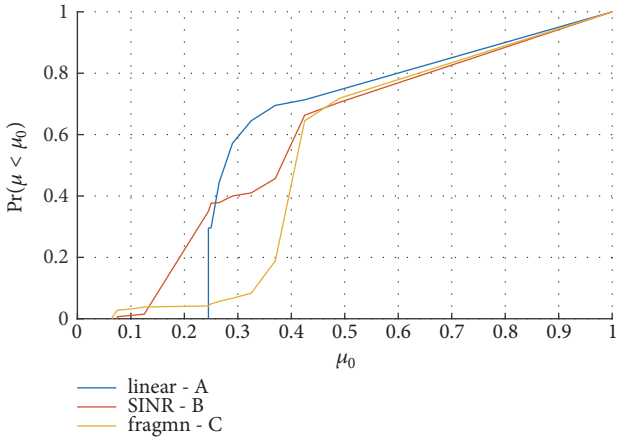


FIGURE 5: CDF of fragmentation coefficient observed at the third base station (only eMBB slice can be deployed).

Table 2: Outage probabilities achieved for three proposed methods, use case 2.

Method	Base station 1	Base station 2	Base station 3	Network	eMBB Slice	mIOT Slice
A	0.1392	0.2975	0.6429	0.3075	0.4405	0.0811
B	0.1456	0.2975	0.6548	0.3125	0.4643	0.0541
C	0.1435	0.3017	0.6310	0.3083	0.4537	0.0608

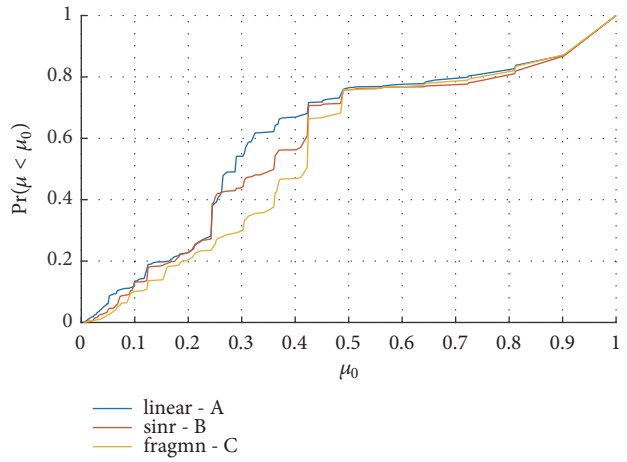


Figure 6: CDF of fragmentation coefficient observed from the network perspective.

observe that, first, in all cases, method C again achieves the best performance in terms of fragmentation and, again for low and high values of μ_0, all methods achieve similar results.

It is also good to analyse the observed outage probability in the network use case. In Table 2 we showed the achieved values in various configurations, mainly, per base station (i.e., the probability that certain base station will be in outage), per slice, and for the entire network. Surprisingly, in a multicell scenario, the achieved values of observed outage probability are almost equal, and the differences between the methods tend to be statistically insignificant. Moreover, the values themselves are high and cannot be accepted in practical situations. This is mainly due to the fact that in the considered use case the algorithm very often cannot allocate the resource regardless of the selected waveform and chosen resource allocation scheme. The mutual constraints between the neighbouring cells and the need for maximisation of the coverage area by a given slice also have a direct impact on the achieved values. Finally, please note that in our solution we have applied a greedy approach, which can be significantly improved.

6.3. Use Case 3: Dominance of the Waveform Selection Phase. In the previous scenarios we have observed the performance of the two-phase algorithm, and one may question the real impact of the waveform selection phase on the final performance. In order to highlight the importance of this step, we have verified the performance of the proposed solution in a scenario with limited resources. The simulation setup applied in this use case is the same as in the first one; however we assume that

(i) there three types of slices: one for mIOT traffic (where the GMSK modulation is applied and which is allocated for time significantly longer than other slices), one slice for eMBB traffic (where only OFDM waveform can be selected), and one for URLLC traffic (where either OFDM or NC-OFDM can be selected).

(ii) for the eMBB slice the appearance of new network slice request was modelled as exponential process with mean value set to 50 time units and with duration also modelled as random variable with exponential distribution with mean value of 50 time units,

(iii) for the URLLC slice the appearance of new network slice request was modelled as exponential process with mean value set to 20 time units and with duration also modelled as random variable with exponential distribution with mean value of 20 time units,

(iv) the SIP offered C=15 channels of 500 kHz of bandwidth

(v) the eMBB and URLLC slices requested 6 channels each (3 MHz in total for each slice) whereas mIOT slice requested just one channel (500 kHz).

(vi) GMSK signal is present in the fifth channel.

One may observe that in such a scenario only six different frequency allocations are possible for the three considered methods, as indicated in Figure 7, where the allocation indexed with 1 represents the initial stage; i.e., there is a long-term running mIOT slice. In the considered scenario, once the eMBB slice is allocated, it is not possible to allocate the new eMBB slice. At the same time, the new URLLC slice, which accepts more advanced waveforms, may be allocated even if there is already one eMBB or URLLC slice. Furthermore, there is a gap between mIOT slice and other OFDM-bases slices due to the need of minimising the interslice interference. We have also assumed that there is no need for a gap between two OFDM slices as discussed in the interference analysis section. One may observe that situations 2, 3, and 4 will be achieved by methods B and C whereas 5, 6, and 7 will be achieved only by method A.

For this situation we have achieved the outage probabilities of 0.4549 and 0.1567 for eMBB and URLLC slices, respectively (of course, the outage probability for mIOT slice is zero). As the number of possible combinations of slice allocation is highly limited, the final performance is affected mainly by the waveform selection phase. In consequence, the achieved outage probability is almost the same for each method (A, B, and C). One may observe that the URLLC slice, which has an option for NC-OFDM waveform, achieves significantly lower outage probability when compared to the

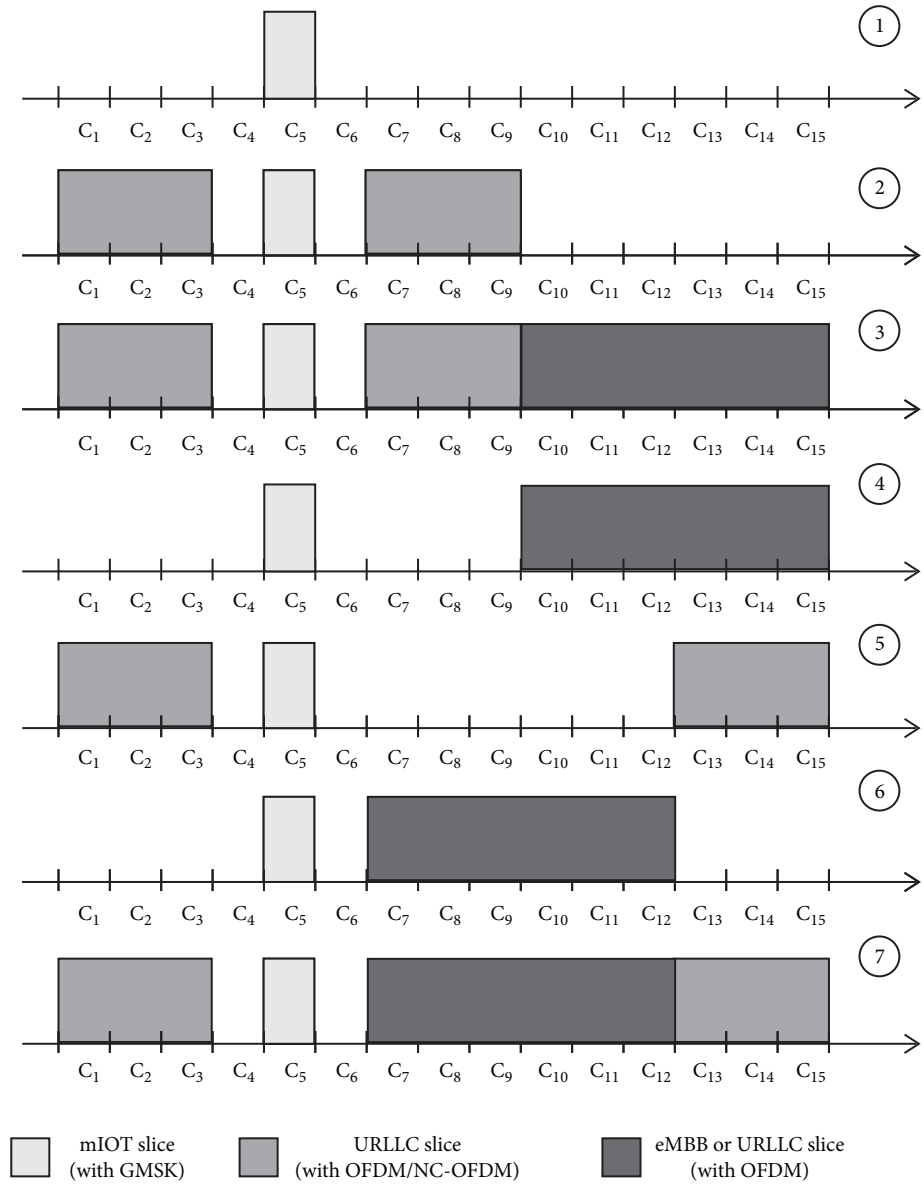


Figure 7: Possible slice allocations.

eMBB slice. This is because the ability to choose noncontiguous waveform creates the opportunity for the assignment of noncontiguous resources. Because there are only six values of the fragmentation coefficient, the CDF of fragmentation coefficient does not provide any new technical insight; in consequence it is not shown here.

6.4. Use Case 4: Single Base Station with Three Slices. Finally, we decided to analyse the performance of the proposed method, when there are three different types of slices available in the single base station, i.e., mIOT, eMBB, and URLLC. In general, the same setup as for use case 1 has been selected; mainly the band $W = 10$ MHz is split into $C = 20$ channels, each of 500 kHz of bandwidth. Moreover, the appearance and duration of each slice was modelled with exponential distribution with different intensities; i.e., eMBB slice (OFDM only, 3 MHz band) has intensity of appearance set to 50 time units (mean value of the exponential distribution) and duration set to 50 time units, URLLC slice (OFDM only, 3 MHz band)–5 and 20 time units, respectively, and mIOT (GMSK, 0.5 MHz band)–20 and 50. Finally, the probability of occurrence of the new slice was set to 0.3 for eMBB and URLLC and 0.4 for mIOT. The results have been achieved over 100000 decisions.

First, let us present the CDF of the fragmentation coefficient (Figure 8). Although the best results are still achieved for method C (especially in the range from 0.3 to 0.5), the differences between these methods are not that significant. We claim that the value of fragmentation coefficient highly depends on the intensities of each slice generation process. Although the differences in the CDF are not that significant, the achieved values of outage probability prove that method

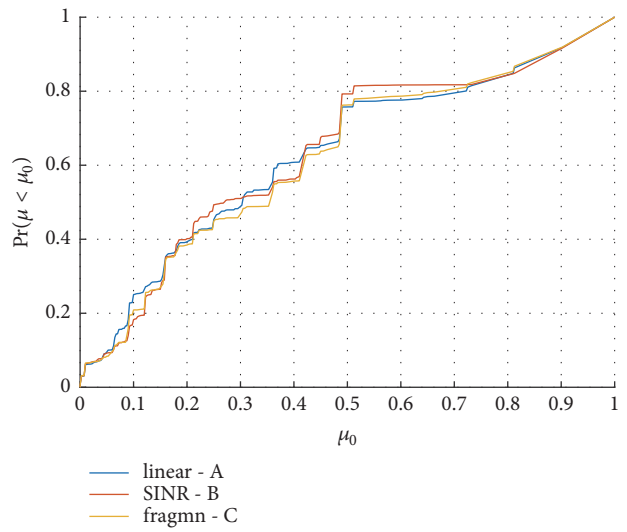


Figure 8: CDF of fragmentation coefficient observed from the network perspective.

Table 3: Achieved outage probabilities achieved for three slices.

Method	Network	mIOT slice	eMBB Slice	URLLC Slice
A	0.1048	0.0047	0.0851	0.2578
B	0.1296	0.0025	0.1102	0.3182
C	0.0955	0.0026	0.0759	0.2387

C (i.e., the one which is optimal from the point of view of the spectrum fragmentation) guarantees the best results, as shown in Table 3. It may be observed that the overall base station performance is mostly dominated by the most frequent slice, i.e., the one that appears with mean value set to 5 time units. As for this slice six channels are requested, it happens quite often that no resources may be allocated to this slice. It shows, however, that such a greedy approach results in unacceptably high outage probability for URLLC slice, where the strict requirements for end-to-end delay are defined. Thus, probably, in the future some kind of priorities shall be defined, and advanced optimisation procedures shall be applied.

7. Summary and Conclusion

In our work we have discussed the possibility of applying the waveform flexibility concept with adaptive spectrum resource assignment to network slicing. The proposed heuristic algorithms for waveform selection and frequency assignment are designed to maximise the revenue of the spectrum and infrastructure provider under the constraint that the assumed quality of service metrics per each slice is guaranteed. Thus, in the first step, which is common for all proposed algorithms, we decided to select such a waveform that, on the one hand, fulfils all requirements specified for the certain network slice (e.g., minimum rate) and, on the other, minimises energy consumption required for data processing. In a nutshell, at all times the simplest possible waveform is selected. Next, once the waveform is selected and the corresponding bandwidth is known, three various approaches to frequency assignment have been tested. The performance of these methods has been verified in four simulation use cases. The obtained results present the performance of these methods indicating that there are ways to optimise the spectrum fragmentation. It has to be stated that the results highly depend on various parameters (such as traffic intensity and definition of the slice in terms of physical and medium access control layers), and thus further investigations toward minimisation of spectrum fragmentation are possible. In a nutshell, however, the results presented in this work prove that there is a possibility of flexible and adaptive resource assignment that would be highly helpful in the practical implementation of network slice concept.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The presented work has been funded by, first, the National Science Centre in Poland within the SONATA Project based on decision no. DEC-2015/17/D/ST7/04078 and, second, by the Polish Ministry of Science and Higher Education within the Status Activity Task (in 2019).

References

[1] J. Mitola III and G. Q. Maguire Jr., "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13–18, 1999.

[2] N. Wang, Y. Gao, and B. Evans, "Database-augmented spectrum sensing algorithm for cognitive radio," in *Proceedings of the IEEE International Conference on Communications, ICC 2015*, pp. 7468–7473, London, UK, June 2015.

[3] Ofcom, "Digital dividend: cognitive access Statement on licence-exempting cognitive devices using interleaved spectrum," 2009, https://www.ofcom.org.uk/__data/assets/pdf_file/0023/40838/statement.pdf.

[4] ECC Report 159, "Technical and operational requirements for the possible operation of cognitive radio systems in the 'white spaces' of the frequency band 470-790 MHz," Cardiff, 2011, https://www.ecodocdb.dk/download/be051b35-91e9/ECCREP159.PDF.

[5] P. Tengkvist, G. P. Koudouridis, C. Qvarfordt, M. Dryjanski, and M. Cellier, "Multi-dimensional radio service maps for position-based self-organized networks," in *Proceedings of the 22nd IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, CAMAD 2017*, pp. 1–6, Sweden, June 2017.

[6] ETSI, "ETSI TS 103 235: system architecture and high level procedures for operation of licensed shared access (LSA) in the 2 300 MHz - 2 400 MHz band," Tech. Rep., 2015.

[7] FCC, "Shared commercial operations in the 3550-3650 MHz band; 47 CFR parts 0, 1, 2, 90, 95, and 96," Tech. Rep., 2015, https://www.law.cornell.edu/rio/citation/80%20FR%2036222.

[8] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: a survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 462–476, 2016.

[9] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: a survey on principles, enabling technologies, and solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[10] J. Perez-Romero, O. Sallent, R. Ferrus, and R. Agusti, "On the configuration of radio resource management in a sliced RAN," in *Proceedings of the NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, pp. 1–6, Taipei, Taiwan, April 2018.

[11] L. Caeiro, F. D. Cardoso, and L. M. Correia, "Addressing multiple virtual resources in the same geographical area," in *Proceedings of the 2015 European Conference on Networks and Communications (EuCNC)*, pp. 185–189, Paris, France, June 2015.

[12] B. Han, J. Lianghai, and H. D. Schotten, "Slice as an evolutionary service: genetic optimization for inter-slice resource management in 5G networks," *IEEE Access*, vol. 6, pp. 33137–33147, 2018.

[13] E. Pateromichelakis and K. Samdanis, "A graph coloring based inter-slice resource management for 5G dynamic TDD RANs," in *Proceedings of the 2018 IEEE International Conference on Communications (ICC 2018)*, pp. 1–6, Kansas City, Mo, USA, May 2018.

[14] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: the 5G network slice broker," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 32–39, 2016.

[15] L. Kulacz, P. Kryszkiewicz, A. Kliks, and J. Perez-Romero, "Waveform flexibility in database-oriented cognitive wireless systems," in *Proceedings of the 2018 Global Information Infrastructure and Networking Symposium (GIIS)*, pp. 1–4, Thessaloniki, Greece, October 2018.

[16] 3GPP, TS 23.501, "Technical specification group services and system aspects; system architecture for the 5G system; stage 2 (release 15)", V15.4.0, 2018.

[17] 3GPP, TS 22.261, "Technical specification group services and system aspects; service requirements for the 5G system; stage 1 (release 16)", V16.6.0, 2018.

[18] GSM Association, "Network slicing, use case requirements," 2018, https://www.gsma.com/futurenetworks/wp-content/uploads/2018/04/NS-Final.pdf.

[19] WWRF, "White paper 3: end to end network slicing," Outlook 21, 2017, https://www.wwrf.ch/files/wwrf/content/files/publications/outlook/White%20Paper%203-End%20to%20End%20Network%20Slicing.pdf.

[20] S. Popli, R. K. Jha, and S. Jain, "A survey on energy efficient narrowband internet of things (NBIoT): architecture, application and challenges," *IEEE Access*, vol. 7, pp. 16739–16776, 2019.

[21] H. Bogucka, P. Kryszkiewicz, and A. Kliks, "Dynamic spectrum aggregation for future 5G communications," *IEEE Communications Magazine*, vol. 53, no. 5, pp. 35–43, 2015.

[22] PHYDYAS, "FBMC physical layer: a primer," Tech. Rep., 2010, http://www.ict-phydyas.org/teamspace/internal-folder/FBMC-Primer_06-2010.pdf.

[23] P. Kryszkiewicz, A. Kliks, and H. Bogucka, "Small-scale spectrum aggregation and sharing," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 10, pp. 2630–2641, 2016.

[24] CEPT Report 40, "Compatibility study for LTE and WiMAX operating within the bands 880-915 MHz / 925-960 MHz and 1710-1785 MHz / 1805-1880 MHz (900/1800 MHz bands)," 2010.

[25] ETSI, "Digital cellular telecommunications system (phase 2+); radio transmission and reception, document ETSI TS 5.05," 1996.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 1745904, 11 pages
https://doi.org/10.1155/2018/1745904

Research Article

A Biological Model for Resource Allocation and User Dynamics in Virtualized HetNet

Lu Ma,[1,2] **Xiangming Wen**,[1,2] **Luhan Wang**,[1,2] **Zhaoming Lu**,[1,2] **Raymond Knopp**,[3] **and Irfan Ghauri**[3]

[1]*Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*Beijing Advanced Innovation Center for Future Internet Technology, Beijing 100124, China*
[3]*Communication System Department, EURECOM, 06410 Biot, France*

Correspondence should be addressed to Lu Ma; malu@bupt.edu.cn

Received 20 April 2018; Revised 14 August 2018; Accepted 9 September 2018; Published 27 September 2018

Guest Editor: Dejan Vukobratovic

Virtualization technology is considered an effective measure to enhance resource utilization and interference management via radio resource abstraction in heterogeneous networks (HetNet). The critical challenge in wireless virtualization is virtual resource allocation on which substantial works have been done. However, most existing researches on virtual resource allocation focus on improving total utility. Different from the existing works, we investigate the dynamic-aware virtual radio resource allocation in virtualization based HetNet considering utility and fairness. A virtual radio resource management framework is proposed, where the radio resources of different physical networks are virtualized into a virtual resource pool and mobile virtual network operators (MVNOs) compete for virtual resources from the pool to provide service to users. A virtual radio resource allocation algorithm based on biological model is developed, considering system utility, fairness, and dynamics. Simulation results are provided to verify that the proposed virtual resource allocation algorithm not only converges within a few iterations, but also achieves a better trade-off between total utility and fairness than existing algorithm. Besides, it can also be utilized to analyze the population dynamics of system.

1. Introduction

Within the last decade, mobile networks are experiencing dramatic increases in data traffic and services [1]. Mobile network operators deploy many kinds of networks and make them denser and denser. For example, in LTE, heterogeneous networks allow a mixed deployment of macro and micro base stations (BSs) in a geographic area providing different accessing capabilities and capacity/coverage needs [2]. However, chaotic and dense deployments cause new problems [3]. First, clients of one BS suffer high interference from neighboring BSs because of the high frequency reuse and broadcast nature of wireless communication. Second, coverage of cellular networks becomes denser and more complex which leads to more users at the edge of network. More handovers and unbalanced load are caused by user mobility. As a result, the benefit in terms of capacity of heterogeneous networks could be undermined if effective measures are not taken [4]. The wireless network virtualization which attracts much attention has recently been considered as a promising solution to increase the spectrum and infrastructure efficiency [5].

With virtualization technology, the wireless network infrastructure can be decoupled from services it provides so that different services can share the same infrastructure [6]. In wireless virtualization, the physical radio resources of heterogeneous networks owned by Infrastructure Providers (InPs) are abstracted and sliced into virtual radio resources and form a virtual radio resource pool [7]. Mobile virtual network operators (MVNOs) lease the virtual radio resources and assign them to users [8]. Since the infrastructure and physical resources of HetNet are abstracted and sliced into virtual resource, many effective measures could be carried out more easily, such as interference management, load balancing, and so on. Furthermore, it is possible that different MVNOs coexist on the same InP and share the infrastructure

and radio resources, which maximizes the utilization of resources and reduces the capital expense (CapEx) and operation expense (OpEx).

In the virtualization environment of HetNet, it is a critical issue to allocate the virtual resources to users in an efficient way. Substantial efforts have been done to research the virtual resource allocation model [8–11]. However, these papers mainly focus on the virtual resource allocation with the assumption that the system is in a steady state; that is, the system has reached its equilibrium. In a system with limited resources, it will take a certain period of time to find the equilibrium sharing the virtual resources into MVNOs to optimize the system performance. As a result, the analysis through steady state based method may not be sufficient which ignores the transient dynamics to reach equilibrium of system. More specifically, it is helpful for making more effective approaches (e.g., interference management, base station sleeping, etc.) to promote system performance if we get the users' behaviors or dynamics. Moreover, most of the existing works aim to maximize the system performance (e.g., utility, throughput, etc.) but do not consider the fairness for MVNOs.

Biological approaches are regarded as an effective method to analyze the time dynamic behaviors of heterogeneous system and some works have been done in this area [12]. In [13], the evolutionary game model was utilized to study the robust equilibrium and the dynamics in wireless networks. However, in [14, 15], the dynamics of cognitive radio networks were studied, respectively, using predator-prey model and evolutionary game model.

However, most of these works mainly focused on analyzing the dynamics and fairness in resource allocation procedure, ignoring the system utility. To the best of our knowledge, the dynamic-aware virtual radio resource allocation for MVNOs in virtualization based HetNet considering utility and fairness has not been studied in previous works.

In this paper, we investigate the virtual radio resource allocation problem for virtualized HetNet based on the biological Lotka-Volterra model [16, 17] with consideration of user dynamics, fairness for MVNOs, and system utility. In the proposed architecture, users of different MVNOs utilize the virtual radio resources in the resource pool in the virtualization based HetNet system. Users of MVNOs benefit from occupying the virtual radio resources and the population of virtual radio resources increases after users release resources. As a result, the relations among different MVNOs and the virtual radio resources are similar to the resource competing of species in a natural environment. The virtual radio resources and users from different MVNOs can be considered as the environment resources and species in natural systems.

The main contributions of this paper are summarized as follows:

(i) We formulate the virtual radio resource allocation problem in virtualization based HetNet as a population competing model, where the users in the system are considered as the predators in nature environment, and virtual radio resources are preys to users. Users of different MVNOs compete for virtual radio resources from the virtual resource pool.

(ii) We introduce the aggregated utility function into the Lotka-Volterra model and take the fairness and utility into consideration at the same time. Furthermore, a virtual resource allocation algorithm based on Lotka-Volterra model is developed in the proposed virtual radio resource management framework.

(iii) The proposed algorithm can quickly capture the time dynamics of system. It is helpful to investigate the behaviors and dynamics of users through the trace of time.

(iv) Simulations are conducted to demonstrate that the proposed virtual resource allocation algorithm outperforms the existing centralized maximal utility (MU) approach, achieving a good trade-off between total utility and fairness.

The rest of this paper is organized as follows. Sections 2 and 3 present the system model and the problem formulation. Section 4 provides the proposed virtual radio resource allocation algorithm in virtualization based HetNet. Analysis of system equilibrium state in the proposed system is presented in Section 5, while in Section 6, performance of the proposed algorithm is evaluated by simulations. Finally, Section 7 concludes the paper.

2. System Model

In this section, we will introduce the wireless virtualization in heterogeneous network and the classic Lotka-Volterra model in ecosystem. We will then propose a model of virtual radio resource allocation based on Lotka-Volterra in HetNet.

2.1. Wireless Virtualization in HetNet. Similar to the network virtualization, wireless virtualization needs physical resources to be abstracted into a number of virtual resources. All the virtual resources are in a virtual resource pool which can be utilized by different service providers. In the virtualization based HetNet system, heterogeneous physical radio resources owned by different Infrastructure Providers (InPs) can be abstracted and sliced into virtual wireless resources which can be shared by multiple mobile virtual network operators (MVNOs). The virtual radio resource management framework in virtualized heterogeneous networks is as shown in Figure 1.

The modules and their functions are described as follows:

(1) InPs: The InPs consist of heterogeneous wireless networks (e.g., macro and small cell base stations). InPs own the physical wireless network infrastructure resources and physical radio resources. They can provide these physical radio resources for MVNOs and get revenue.

(2) Hypervisor: The hypervisor virtualizes the physical resources from different InPs and enables the sharing for MVNOs. The allocation and mapping of the virtual resources are realized by Virtual Resource Management and Virtual Resource Mapping modules. Also, the hypervisor is responsible for collecting the users' information from MVNOs.

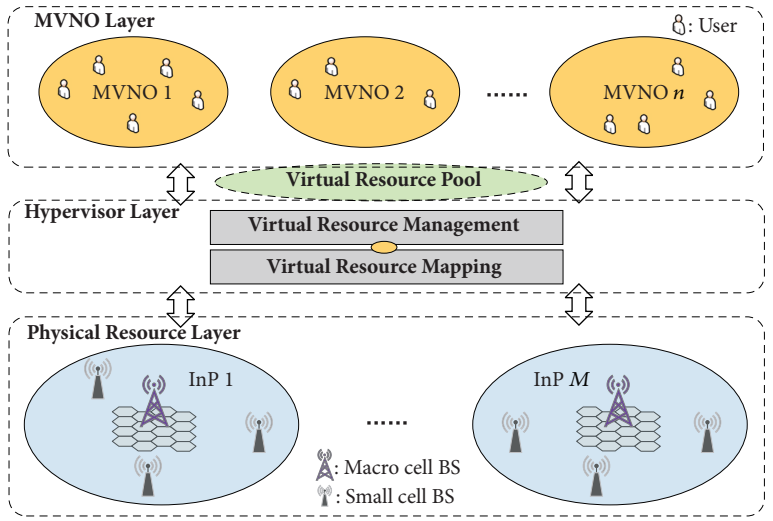


Figure 1: The virtual radio resource management framework in virtualized HetNet.

(3) MVNOs: MVNOs lease the physical radio resources from InPs, abstract them into virtual radio resources based on the requests from users, and assign the virtual radio resources to each user. As a result, the MVNOs can provide various services to their subscribers through the same substrate networks.

The virtualization technology can enable the resource sharing in heterogeneous wireless networks which will reduce the CapEx and OpEx. The authors of [18] have estimated that 40 percent of 60 billion USD may be saved using the virtualization technology in wireless networks. In the virtualization based heterogeneous networks, a significant issue is how to catch the time dynamics of users and allocate virtual radio resources into multiple MVNOs which is the purpose of this paper.

2.2. The Lotka-Volterra Model. The Lotka-Volterra model is a mathematical population model of biology which was developed by Alfred J. Lotka (1925) [19] and Vito Volterra (1926) [20]. Classic Lotka-Volterra competition model describes relationship and dynamics of different animal populations competing for shared resources. Assuming that x_1 and x_2 represent populations of two species and all the parameters in this model are positive, then

$$\begin{aligned} \dot{x}_1 &= r_1 x_1 \left(1 - \frac{x_1}{K_1} - x_2 \frac{\alpha_1}{K_1}\right) \\ \dot{x}_2 &= r_2 x_2 \left(1 - \frac{x_2}{K_2} - x_1 \frac{\alpha_2}{K_2}\right) \end{aligned} \tag{1}$$

The parameters r_1 and r_2, respectively, refer to the intrinsic growth rate of the two species, K_1 and K_2 are the carrying capacity of the two species, and α_1 and α_2 are the inter-specific competition coefficient which reflects that the individuals of one species have inhibited influence on the competitors of the other species. In this model, for example, α_1 denotes the inhibited effect the individual of species x_1 has on x_2, that is, the resource each individual of species x_2 occupies equal to that α_1 individual of species x_1 occupies.

In the Lotka-Volterra model, the population of a species not only depends on the limitation of resources in ecosystem, but also is affected by the survival competing with another species. When a species grows fast, the population of another species will decrease. In the extreme situation, one species can reach its carrying capacity, while the population of another species keeps on the lowest level.

2.3. Proposed Virtual Resource Allocation Model. The resource competing among different species in the natural environment is similar to the virtual radio resource competing in the virtualized wireless networks. In the virtualization based HetNet system, users of different MVNOs utilize the virtual radio resources in the resource pool. Users of MVNOs benefit from occupying the virtual radio resources and the population of virtual radio resources increases after users release resources. Similarly, the population of a MVNO will be limited when other MVNOs occupy too much radio resources (because the radio resources are always limited in wireless system).

Inspired by biological systems and models (i.e., Lotka-Volterra model) in nature, we propose an ecology based model for virtual radio resource allocation and users dynamics analysis in HetNet system.

The proposed ecological model in Figure 2 describes the paradigm of virtual resource competing in HetNet, which consists of preys in the environment and n species. The species represent different mobile virtual network operators (MVNOs), while the environment resources are considered as virtual radio resource pool of HetNet. The MVNOs fed on virtual radio resource blocks must continually evolve to ensure sustainability and meet the changing environment, which is analogous to species that have to survive and evolve by consuming environment resources. Users of every MVNO and the virtual radio resources form a "food chain." Just like the biological system, MVNOs and virtual radio resources

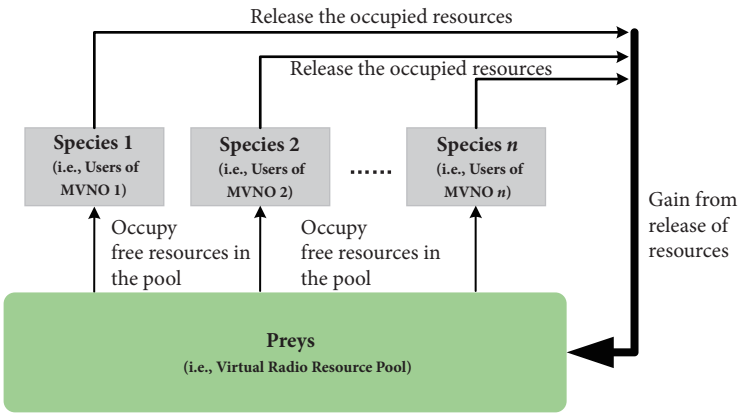


Figure 2: The proposed virtual resource allocation model.

can be considered as predators and preys, respectively, the populations of MVNOs benefit from competing to occupy virtual radio resources, and virtual radio resource pool increases when users release the resources after completing communication. However, the total amount of resources in a communication system is limited and fixed, which is different from the common biological system.

Under specific assumptions, the proposed model can be shown to mechanistically represent competition for resources among species. This can be extended into wireless communication system where MVNOs compete for virtual radio resources.

3. Problem Formulation

In this section, we formulate the virtual resource allocation problem in virtualized HetNet with the biological model. Firstly, we present the utility functions for users and MVNOs. Then, the virtual radio resource allocation is formulated to resource competing among MVNOs according to the utilities of them.

3.1. Utility Function Definition. In our virtualized wireless networks, users lease virtual resources from their MVNOs. The virtual resources allocated to users are mapped to substrate physical networks. Users pay to their MVNOs based on the data rate they can get. Ensuring the total utility of the system is an important goal of this paper, so that the utility function should firstly be defined. By mapping the utility function with the competing coefficient of biological competing model, we could guarantee that the MVNO with higher aggregated utility can get more resources. As a function of data rate, the utility of user j at time t can be mathematically expressed by [21]:

$$u_{j,t} = \frac{r_{j,t}^{1-\beta_j}}{1-\beta_j} \quad \left(0 < \beta_j < 1\right) \tag{2}$$

where $r_{j,t}$ is the potential data rate of user's virtual resource request at time t and β_j represents the traffic type of user request j. The $r_{j,t}$ can be derived by the Shannon formula based on the bandwidth (denoted as $b_{j,t}$) required by the user request j and the effective signal-to-interference-plus-noise ratio at time t (denoted as $SINR_{j,t}$):

$$r_{j,t} = b_{j,t} \cdot \log_2\left(1 + SINR_{j,t}\right) \tag{3}$$

We assume that utilities of virtual resource requests are positive and virtual radio resources are allocated periodically every T time so that time index t can be ignored. The utility function of user j can be reformulated as

$$u_j = \frac{\left[b_j \cdot \log_2\left(1 + SINR_j\right)\right]^{1-\beta_j}}{1-\beta_j} \quad \left(0 < \beta_j < 1, b_j \geq 0\right) \tag{4}$$

Then, we can obtain the aggregated utility of MVNO i,

$$U_i = \sum_j u_j = \sum_j \frac{\left[b_j \cdot \log_2\left(1 + SINR_j\right)\right]^{1-\beta_j}}{1-\beta_j} \tag{5}$$

subject to

$$\sum_j b_j \leq B_{total}, \quad 0 < \beta_j < 1, \quad b_j \geq 0 \tag{6}$$

where B_{total} is the total bandwidth of the system.

3.2. Proposed Virtual Resource Allocation Problem Formulation. In the proposed virtualized wireless HetNet, MVNOs need to compete for virtual resources according to the requests of their users. The objective of this paper is to develop a virtual resource allocation algorithm improving the total utility of system and catching the time dynamics of users which is important for network management.

The population model of every MVNO can be expressed by a system of Ordinary Differential Equations (ODEs) which are powerful tools for modelling dynamic systems that

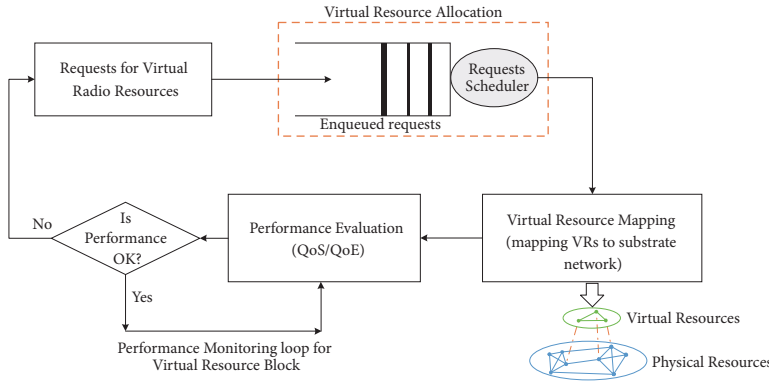


Figure 3: The virtual radio resource allocation process in HetNet.

change with time (d/dt). The population (denoted as $N_i(t)$) is defined as the number of users who are using the virtual radio resources without interruption in MVNO i at time t. The model can be expressed as

$$\dot{N}_i = g_i N_i \left[1 - \frac{1}{K_i} N_i - \frac{\alpha_i}{K_i} \left(\sum_{p=1, p \neq i}^{n} N_p \right) \right] \tag{7}$$

subject to

$$\begin{gathered} g_i > 0, \alpha_i > 0 \\ N_i, K_i \in \mathbb{N}^+ \\ N_i \leq K_i \end{gathered} \tag{8}$$

where g_i is the intrinsic growth rate of users occupying the virtual radio resource, while K_i represents carrying capacity of the system in terms of MVNO i. In this model, α_i represents the average competition coefficient of other MVNOs on MVNO i, and the term $\alpha_i \left(\sum_{p=1, p \neq i}^{n} N_p\right)$ can be thought of as the decrease in growth rate of MVNO i due to the presence of other MVNOs, so that N_i/K_i and $(\alpha_i/K_i)(\sum_{p=1, p \neq i}^{n} N_p)$ can be, respectively, seen as the virtual radio resources already occupied by MVNO i and other MVNOs.

To ensure utility of the whole system, we define the competition coefficient of virtual network i associated with the aggregated utility. Firstly, we assume that the competing coefficient α_i is inversely associated with the aggregated utility U_i so that the MVNO with higher aggregated utility will get more resources. Then, we introduce the adjustment factor a_i to make sure that the MVNO with higher priority can get lower competing coefficient, and as a result the priorities of some MVNOs can be guaranteed. So, we define the competing coefficient of virtual network i as $\alpha_i = a_i + c \cdot (1/U_i)$ (c is a constant to make sure the values of α_i satisfy the conditions of convergence). Now we can derive the differential equations of population density as

$$\dot{N}_i = g_i N_i \left\{ 1 - \frac{1}{K_i} N_i - \frac{a_i + c \sum_j \left(\left(1 - \beta_j\right) / \left[b_j \cdot \log_2 \left(1 + SINR_j\right) \right]^{1-\beta_j} \right)}{K_i} \left(\sum_{p=1, p \neq i}^{n} N_p \right) \right\} \tag{9}$$

The MVNO with bigger potential aggregated utility may occupy more virtual resource block because it has lower average competition rate with other populations. In each MVNO, virtual resource blocks in the resource pool are assigned to users according to their potential utility u_j. And generally, fairness and priorities of some virtual networks can be guaranteed via adjusting the variable a_i. On the one hand, we can set a minimum value of a_i for every virtual network so that the coexistence and fairness are maintained. On the other hand, we can set a bigger value for a MVNO with priority to guarantee that it can occupy more resources even when it has lower utility than others.

4. The Virtual Radio Resource Allocation Algorithm

In this section, we first present the virtual radio resource allocation progress in virtualization based HetNet and then propose a virtual radio resource allocation algorithm based on the biological competing model.

The system-theoretical model for virtual radio resource allocation progress is as shown in Figure 3. Users request virtual radio resources from their MVNOs and get corresponding services. The requests for virtual radio resources arrive in real time. Then MVNO will conduct the virtual resource allocation for its users. Firstly, the requests will be enqueued at MVNO's master queue according to their priorities and arrival time. Then they will be scheduled to occupy the resources every T time based on their requirement and service level agreement. This procedure is called virtual resource allocation which is investigated in this paper. Then, the virtual resource mapping is conducted, and after the virtual resources are mapped to substrate physical networks, users can start the data transmission procedures. Closed performance monitoring loop can be adopted to evaluate

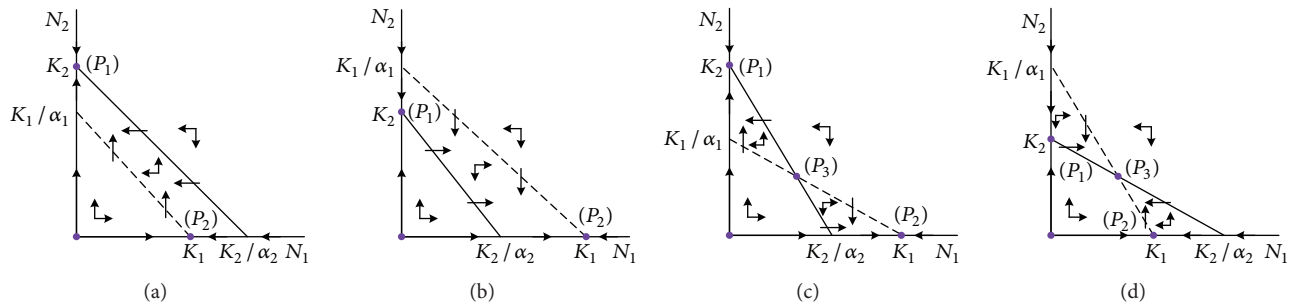


FIGURE 4: The isocline cases with steady states and flow patterns for the proposed model. Thick dashed lines indicate N_1-isoclines; thick solid lines represent N_2-isoclines.

the performance dynamically. If the performance could not satisfy user's requirement (because the physical channel is changing dynamically), this request will be scheduled in the next period of time.

The proposed model is then applied to virtual radio resource allocation in heterogeneous wireless networks. In the proposed model, the MVNO has bigger value of α and can occupy less resources because other populations have more inhibited influence on it. The model can guarantee the coexistence of all MVNOs with stability and fairness. Thus we can dynamically allocate the virtual radio resources in the pool of HetNet by adjusting the value of α and maintaining the competing and coexisting relationship among MVNOs. We assume that all resource allocation is synchronously performed every T time. Each MVNO gets virtual resources based on its competition coefficient α_i which is a function of aggregated utility of MVNO i. MVNOs are able to manage the virtual resources and allocate them to users according to their aggregated utility and users' demands. The competition coefficient of each MVNO can also be adjusted based on the knowledge and user dynamics of the whole system to get a more efficient utilizing of resources. The details of virtual radio resource allocation algorithm are described as in Algorithm 1.

5. Analysis of System Equilibrium State in the Proposed System

In this section, we present the analysis of system equilibrium state from the aspects of existence, type, and stability.

5.1. The Existence and Category of Equilibrium Points. The important issue regarding system dynamics is the equilibrium of the model. In order to simplify the problem, let us discuss the system which consists of two MVNOs. We could get an algebraic equation set with the right sides of the OEDs (7):

$$\begin{aligned} g_1 N_1 \left(1 - \frac{1}{K_1} N_1 - \frac{\alpha_1}{K_1} N_2\right) &= 0 \\ g_2 N_2 \left(1 - \frac{1}{K_2} N_2 - \frac{\alpha_2}{K_2} N_1\right) &= 0 \end{aligned} \tag{10}$$

As known in mathematics, the real roots of this algebraic equation set are called the equilibrium point of the model. For equilibrium point $P_0({N_1}^0, {N_2}^0)$, if the solutions of ODEs always meet the restrictions $\lim_{t\to\infty} N_1(t) = {N_1}^0$ and $\lim_{t\to\infty} N_2(t) = {N_2}^0$ at any initial condition, we call P_0 a stable equilibrium point (asymptotic stability); otherwise P_0 is unstable (not asymptotic stability). Obviously, there are two equilibrium points $P_1 = (0, K_2)$ and $P_2 = (K_1, 0)$ in addition to the trivial equilibrium point $N_1 = N_2 = 0$ for the set of ODEs (7) ($n = 2$). When the algebraic equation set (13) has a nonnegative solution $N_1 = {N_1}^*$, $N_2 = {N_2}^*$, we find the third equilibrium point of the system, that is, $P_3({N_1}^*, {N_2}^*)$.

To analyze the properties of this system, phase plane analysis of the model equations is carried out. Firstly, we plot the zero growth rate line for each species taking N_1 and N_2 (number of users occupying resources in MVNOs 1 and 2) as the coordinate. The zero growth line can be obtained by drawing a line between N_1 and N_2 intercepts of each species. Users in both species will increase until they reach the zero growth isocline. The MVNO represented on the abscissa meets the isocline horizontally, and the one represented on the ordinate will do it vertically. Figure 4 illustrates the flow patterns and four possible types of relations between the isoclines of MVNOs 1 and 2. From the flows across the isoclines, the direction fields within each separate region in the phase plane can be worked out. For example, in case (a), it is obvious within the area bounded by these two isoclines that the flow is towards the top-left corner, which implies that all solutions will tend to the N_2-only state $(N_1, N_2) = (0, K_2)$.

Hence, we can find three equilibrium points for the proposed model (excluding the trivial equilibrium point): $P_1 = (0, K_2)$, $P_2 = (K_1, 0)$, $P_3 = ({N_1}^*, {N_2}^*) = ((K_1 - \alpha_1 K_2)/(1 - \alpha_1\alpha_2), (K_2 - \alpha_2 K_1)/(1 - \alpha_1\alpha_2))$. In the N_2-only state, species 2 reaches its carrying capacity and species 1 goes extinct, while in the N_1-only state, species 2 goes extinct and species 1 tends to its carrying capacity. In cases (c) and (d), there is a coexistence state in which both species have nonzero abundance.

5.2. The Stability of Equilibrium Points. Now, we analyze the stability of equilibrium points. Since the proposed model is applied to wireless communication system and the users of each MVNO cannot be decreased to zero considering the revenue of MVNOs, we need not to consider equilibrium points $P_1 = (0, K_2)$ and $P_2 = (K_1, 0)$. Furthermore, wireless communication systems should be guaranteed the asymptotic

```
1: for Each MVNO i do
2:     for Each user j do
3:         Check the state of data transmission every T time
4:         if Data transmission has been accomplished then
5:             Release the resources occupied by user j
6:         end if
7:     end for
8:     Update the carrying capacity K_i every T time
9:     if K_i(T) ≠ K_i(T − 1) then
10:        Update K_i
11:    end if
12:    Update competition coefficient α_i every T time according to (5)
13:    if α_i(T) ≠ α_i(T − 1) then
14:        Update α_i
15:    end if
16:    Allocate virtual radio resources to MVNO i according to Equations (9)
17:    Sort the users' resource requests of MVNO i by their potential utility u_j
18:    for Each user with the largest utility do
19:        Assign the requested virtual radio resource blocks that satisfy restrictions to it
20:        if Failing in Line 19 then
21:            Reject the request and postpone it to the waiting queue, Break
22:        end if
23:        Users that have occupied resource can begin data transmission
24:    end for
25: end for
```

Algorithm 1: Biological Competing Resource Allocation Algorithm.

stability, so that it is necessary to make sure whether the equilibrium point $P_3(N_1^*, N_2^*)$ is stable. Firstly, we give a proposition about the stability of equilibrium points of (7) ($n = 2$).

Proposition 1. *If the two inequalities $K_2 < K_1/\alpha_1$, $K_1 < K_2/\alpha_2$ are satisfied, the system has three equilibrium points $P_1 = (0, K_2)$, $P_2 = (K_1, 0)$, $P_3(N_1^*, N_2^*)$, and $P_3(N_1^*, N_2^*)$ is the stable equilibrium point. The two MVNOs could coexist in the course of time, $N_1(t) \longrightarrow N_1^*$, $N_2(t) \longrightarrow N_2^*$.*

Proof. Because $P_3(N_1^*, N_2^*)$ is the equilibrium point, the linearization of the ODEs (7) ($n = 2$) near point P_3 can be expressed as

$$\begin{aligned}
&\frac{d\left(N_1 - N_1^*\right)}{dt} \\
&\quad = g_1 N_1^* \left(-\frac{1}{K_1}\left(N_1 - N_1^*\right) - \frac{\alpha_1}{K_1}\left(N_2 - N_2^*\right)\right) \\
&\frac{d\left(N_2 - N_2^*\right)}{dt} \\
&\quad = g_2 N_2^* \left(-\frac{1}{K_2}\left(N_1 - N_1^*\right) - \frac{\alpha_2}{K_2}\left(N_2 - N_2^*\right)\right)
\end{aligned} \tag{11}$$

Then, the characteristic equation of the coefficient matrix is

$$\begin{vmatrix} \lambda + \dfrac{g_1}{K_1} N_1^* & \dfrac{g_1 \alpha_1}{K_1} N_1^* \\ \dfrac{g_2 \alpha_2}{K_2} N_2^* & \lambda + \dfrac{g_2}{K_2} N_2^* \end{vmatrix} = 0 \tag{12}$$

That is,

$$\lambda^2 + \left(\frac{g_1}{K_1} N_1^* + \frac{g_2}{K_2} N_2^*\right)\lambda + \Delta N_1^* N_2^* = 0 \tag{13}$$

where $\Delta = g_1 g_2 (1 - \alpha_1 \alpha_2)/K_1 K_2$.

The discriminant of (13) is

$$\begin{aligned}
&\left(\frac{g_1}{K_1} N_1^* + \frac{g_2}{K_2} N_2^*\right)^2 - 4\Delta N_1^* N_2^* \\
&\quad = \left(\frac{g_1}{K_1} N_1^* - \frac{g_2}{K_2} N_2^*\right)^2 + 4\Delta N_1^* N_2^* > 0
\end{aligned} \tag{14}$$

So (13) has two different real eigenvalues: $\lambda_1 < \lambda_2$, and by using the Vieta theorem, we can get

$$\begin{aligned}
\lambda_1 + \lambda_2 &= -\left(\frac{g_1}{K_1} N_1^* + \frac{g_2}{K_2} N_2^*\right) \\
\lambda_1 \lambda_2 &= \Delta N_1^* N_2^*
\end{aligned} \tag{15}$$

because $K_2 < K_1/\alpha_1$ and $K_1 < K_2/\alpha_2$ are satisfied, just as the situation in Figure 4(d), that is, $g_2 K_1/g_1 K_2 \alpha_1 > g_2/g_1 > g_2 K_1 \alpha_2/g_1 K_2$. As a result, $g_2 K_1/g_1 K_2 \alpha_1 - g_2 K_1 \alpha_2/g_1 K_2 = g_2 K_1 (1 - \alpha_1 \alpha_2)/g_1 K_2 \alpha_1 > 0$, so that $\Delta > 0$. We can get that $\lambda_1 < 0$ and $\lambda_1 < 0$ by putting $\Delta > 0$ into (15). Thus, the equilibrium point $P_3(N_1^*, N_2^*)$ is stable (asymptotic stability).

Similarly, we can prove that $P_3(N_1^*, N_2^*)$ is not stable when $K_2 > K_1/\alpha_1$ and $K_1 > K_2/\alpha_2$ are satisfied which is illustrated in Figure 4(c). Actually, in case (c), MVNO

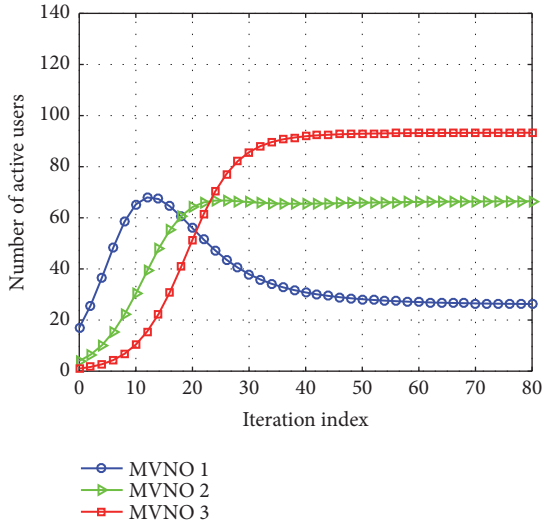


Figure 5: The population dynamics of MVNOs in the proposed system (n=3).

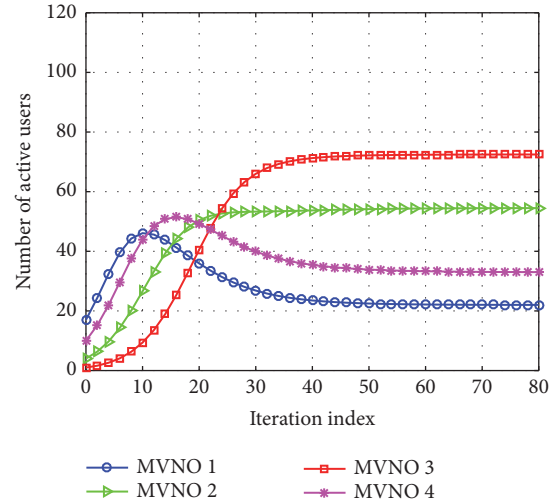


Figure 6: The population dynamics of MVNOs in the proposed system (n=4).

1 can outcompete MVNO 2, but MVNO 2 can also outcompete MVNO 1. The equilibrium point P_3 may evolve to P_1 or P_2 which depends on the initial conditions of the system.

Hence, we could guarantee the coexistence of MVNOs over time by adjusting the values of system parameters every T time based on the proposition and make sure that the proposed system can reach a stable equilibrium. □

6. Simulation Results and Discussion

In this section, simulation results are given to evaluate the performance of the proposed algorithm. In the simulation, the parameters are designed to model a high-loaded system where the available resources are not enough to serve all the users. We assume that there are three MVNOs in the system, and the virtual resource blocks for different user requests in MVNOs have the different bandwidths, which are with uniform distribution. The following parameters are used: $a_1 = 0.08$, $a_2 = 0.05$, $a_3 = 0.03$, $c = 1.6 \times 10^5$, $\beta_j = 0.5$, $b_j \sim U(0, 4)$ MHz, $SINR_j \sim U(5, 20)$ dB. The three MVNOs are assumed to have the different number of users (J) at time t.

Figure 5 illustrates the population dynamics for the proposed system which has three MVNOs (n=3). The convergence of Algorithm 1 is evaluated with resource growth rates $g_i = 0.3$, number of users in each MVNO $J_1 = 100$, $J_2 = 110$, $J_3 = 120$, and carrying capacities $K_1 = 90$, $K_2 = 100$, $K_3 = 110$. As can be seen in Figure 5, the number of active users in each MVNO converges within 50 iterations. This result, together with the previous analysis, indicates that the proposed Algorithm 1 converges to a stable equilibrium.

It also can be seen from Figure 5 that the MVNO with higher service level (smaller a_i) can get more virtual resources to serve its users, because a smaller a_i reduces the competition coefficient, as defined in (9). The MVNO with larger competition coefficient can also get some resources, although the resources are limited in the system. MVNOs coexist in the proposed system and maintain a dynamic balance actually.

In Figure 6, the population dynamics is simulated with four MVNOs to evaluate the convergence of the proposed algorithm when the number of MVNOs scales. As can be seen from Figure 6, the number of active users in each MVNO also converges after iterations. However, the number of MVNOs affects the rate of convergence; more MVNOs may lead to lower convergence rate. For example, the number of active users in MVNO 3 converges within 30 iterations when there are 3 MVNOs in the system, but when there are 4 MVNOs in the system, it converges after 45 iterations. Furthermore, the number of active users of each MVNO will converge to lower level when the system has more MVNOs, because the resources are limited.

As the maximal utility (MU) approach can achieve maximal system utility, it has become a very popular method in virtual resource allocation. For example, the authors in [8] have worked on the virtual resource allocation for MVNOs based on the MU of system. There are also some existing works which focus on the fairness and use the round-robin (RR) method [22]. Therefore, we compare the performance of the proposed Lotka-Volterra solution (LVS) with that of the existing MU approach and RR method.

In Figures 7–9, we compare the utilities of MVNOs when the number of users requesting resource in MVNOs increases from (50, 60, 70) to (100, 110, 120), with resource growth rates $g_i = 0.3$, carrying capacities $K_1 = 90$, $K_2 = 100$, $K_3 = 110$.

As can be seen in Figures 7–9, a MVNO with higher service level gets higher utility in MU approach, RR method, and the proposed Algorithm 1. With MU approach, the utility of MVNO 3 increases linearly while the utility of MVNO 1 drops to zero when the users in MVNO 1 are over 80. However, with the proposed algorithm or RR method, the utilities are always positive because each MVNO can get some resources to serve users. Therefore, it is verified that the proposed algorithm and RR method can guarantee

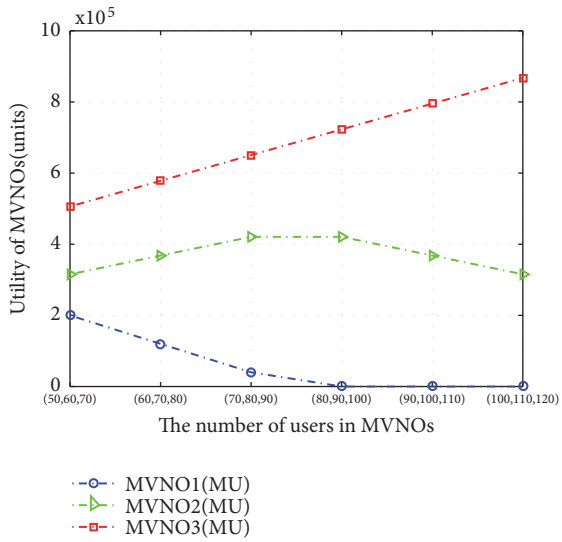


Figure 7: Utility of MVNOs in MU approach.

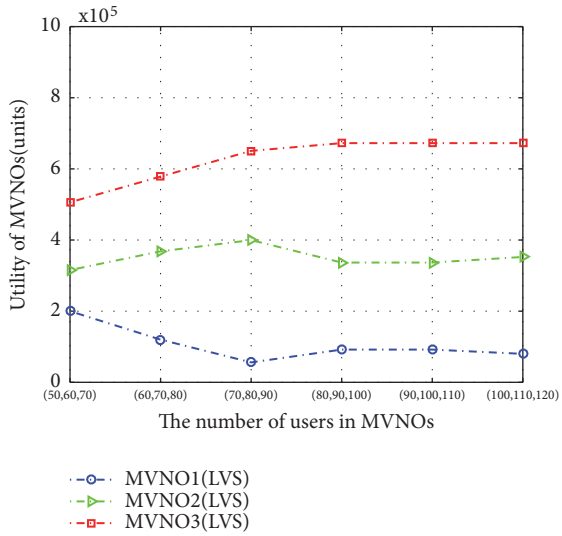


Figure 8: Utility of MVNOs in the proposed Algorithm 1.

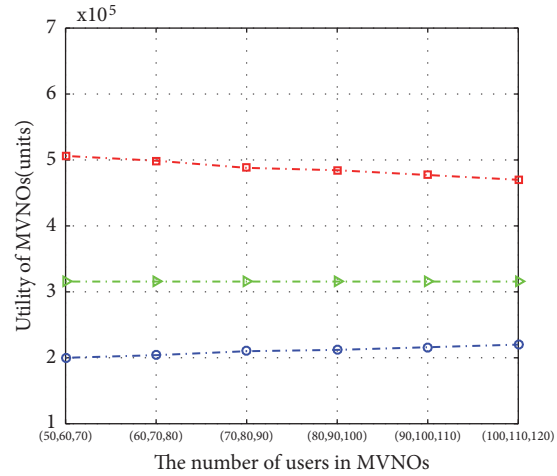


Figure 9: Utility of MVNOs in RR method.

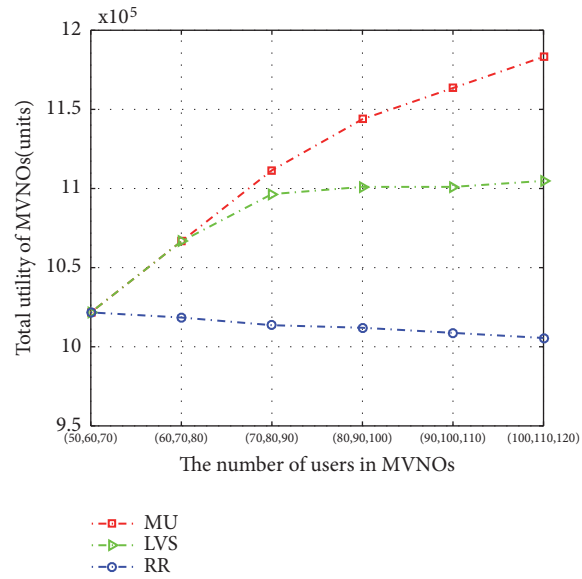


Figure 10: Comparison of the total utilities of different approaches.

higher utility for MVNO with lower service level than the MU approach. Also, they improve the fairness in resource allocation process of the system.

To evaluate the fairness of MVNOs in the system, we use the fairness index (FI), which is defined as [23]:

$$\frac{\left(\sum_{i=1}^{n} N_i\right)^2}{\left(n\left(\sum_{i=1}^{n} \left(N_i\right)^2\right)\right)} \tag{16}$$

The fairness index is widely applied in the literature to evaluate the level of fairness achieved by resource allocation algorithms.

In Figures 10 and 11, we compare the total utility of system and fairness among MVNOs (n=3). As can be seen from Figure 10, the proposed Algorithm 1 achieves higher total utility than RR method, but lower total utility than MU approach when the users in MVNOs are more than (70, 80, 90). Figure 11 shows that the proposed Algorithm 1 achieves lower fairness index than RR method, but higher fairness index than MU approach when the users in MVNOs are more than (70, 80, 90). As one can conclude from Figures 10 and 11, the proposed Algorithm 1 can achieve higher fairness index with the cost of small reduction in total utility of system, compared with the MU approach. It achieves a better trade-off between total utility and fairness than the existing methods.

Besides, we could also analyze the time dynamics of users and virtual radio resources in the system by utilizing the Lotka-Volterra model. The users in the system can be considered as the predators in nature environment, and virtual radio resources are preys to users. The number (i.e.,

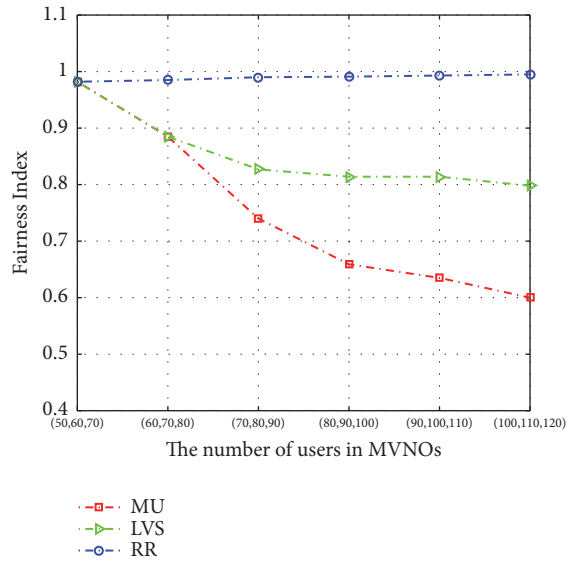


Figure 11: Comparison of the fairness index of different approaches.

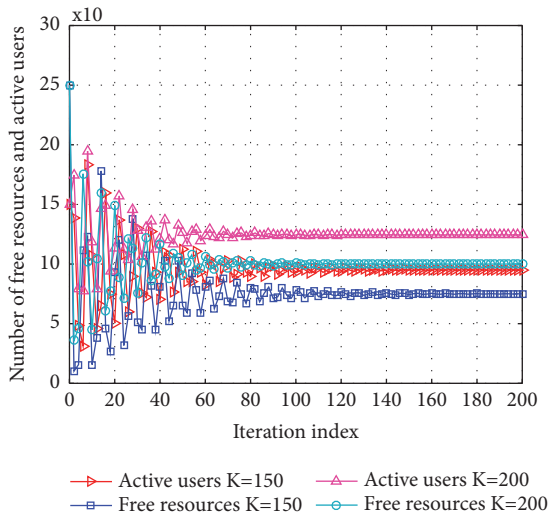


Figure 12: The population dynamics of users and virtual radio resource pool in the proposed system. The parameter K is the system capacity.

population) of active users in the system is benefited by the loss of the population of the virtual radio resources. Therefore, the ecological model is helpful to investigate the behaviors and user dynamics of system through the trace of time. The numerical illustrations of the user dynamics are presented in Figures 12 and 13. From Figure 12, we observe that the behaviors of users in the proposed system follow those in the nature environment; that is, the population of resource will decrease when there are too many users in system. The users and virtual radio resources in the system are regulating to each other. Time dynamics of users in different system capacity (i.e., $K = 150, 200$) are simulated, and by adjusting the system capacity K within certain range, we can get an improvement of system throughput, as shown in Figure 12.

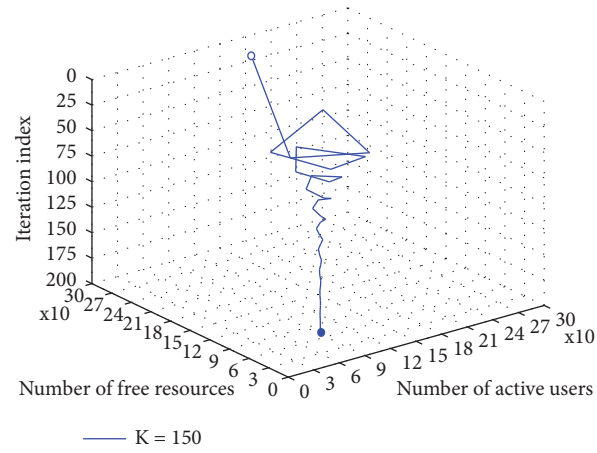


Figure 13: The 3-D phase diagram of the system where users and virtual radio resources are considered predators and preys, respectively.

The phase diagram of user dynamics is illustrated in Figure 13 in which we could catch the dynamics of interaction between users and virtual resources in the system. In this simulation, the system capacity is set as $K = 150$. The system converges to its equilibrium after a period of time. As can be seen in these two figures, the dynamics of users and virtual resources can be analyzed utilizing the proposed biological model, and the throughput of system can be improved by increasing the capacity of system.

By understanding the time dynamics of users, we can further develop effective strategies to enhance system performance. For instance, the hypervisor can conduct adaption of system parameters and base station sleeping strategies to improve the channel utilization and energy efficiency.

7. Conclusion

In this paper, we have investigated the virtual resource allocation problem in virtualization based heterogeneous wireless networks. The virtual resource allocation problem in HetNet was formulated as a population competing model, where the users in the system are considered as the predators in nature environment, and virtual radio resources are preys to users. Users of different MVNOs compete for virtual radio resources from the virtual resource pool. Accordingly, a virtual resource allocation algorithm based on Lotka-Volterra model was developed, considering system utility, fairness, and dynamics. Simulation results showed that the proposed virtual resource allocation algorithm not only converges within a few iterations, but also achieves a better trade-off between total utility and fairness than existing algorithm. Besides, it can also be utilized to analyze the time dynamics of users which is helpful for making more effective approaches to promote system performance.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by China Ministry of Education-CMCC Research Fund Project no. MCM20160104, National Science and Technology Major Project no. 2018ZX030110004, Beijing Municipal Science and Technology Commission Research Fund Project no. Z171100005217001, and Fundamental Research Funds for Central Universities no. 2018RC06. Besides, we would like to thank EURECOM and OpenAirInterface Software Alliance for their support and help.

References

[1] C. Liang and F. R. Yu, "Wireless network virtualization: a survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 358–380, 2015.

[2] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-Dense Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2522–2545, 2016.

[3] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "Softran: software defined radio access network," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking (HotSDN '13)*, pp. 25–30, August 2013.

[4] Q. Li, R. Q. Hu, Y. Qian, and G. Wu, "Intracell cooperation and resource allocation in a heterogeneous network with relays," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 4, pp. 1770–1784, 2013.

[5] C. Liang and F. R. Yu, "Wireless virtualization for next generation mobile cellular networks," *IEEE Wireless Communications Magazine*, vol. 22, no. 1, pp. 61–69, 2015.

[6] A. Blenk, A. Basta, M. Reisslein, and W. Kellerer, "Survey on network virtualization hypervisors for software defined networking," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 655–685, 2016.

[7] S. Fan, H. Tian, and W. Wang, "A Radio Resource Virtualization-Based RAT Selection Scheme in Heterogeneous Networks," *IEEE Communications Letters*, vol. 21, no. 5, pp. 1147–1150, 2017.

[8] C. Liang, F. R. Yu, H. Yao, and Z. Han, "Virtual Resource Allocation in Information-Centric Wireless Networks with Virtualization," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9902–9914, 2016.

[9] T. Leanh, N. H. Tran, D. T. Ngo, and C. S. Hong, "Resource Allocation for Virtualized Wireless Networks with Backhaul Constraints," *IEEE Communications Letters*, vol. 21, no. 1, pp. 148–151, 2017.

[10] W. H. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wireless Communications Magazine*, vol. 21, no. 2, pp. 106–112, 2014.

[11] D. Kreutz, F. M. V. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: a comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015.

[12] S. Balasubramaniam, K. Leibnitz, P. Lio, D. Botvich, and M. Murata, "Biological principles for future Internet architecture design," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 44–52, 2011.

[13] H. Tembine, E. Altman, R. El-Azouzi, and Y. Hayel, "Evolutionary games in wireless networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 3, pp. 634–646, 2010.

[14] D. Liau, K.-C. Chen, and S.-M. Cheng, "A predator-prey model for dynamics of cognitive radios," *IEEE Communications Letters*, vol. 17, no. 3, pp. 467–470, 2013.

[15] S.-M. Cheng, P.-Y. Chen, and K.-C. Chen, "Ecology of cognitive radio Ad hoc networks," *IEEE Communications Letters*, vol. 15, no. 7, pp. 764–766, 2011.

[16] O. T. Solbrig and D. J. Solbrig, *Introduction to Population Ecology and Evolution*, Addison-Wesley, 1979.

[17] S. B. Hsu, S. P. Hubbell, and P. Waltman, "Competing predators," *SIAM Journal on Applied Mathematics*, vol. 35, no. 4, pp. 617–625, 1978.

[18] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 27–35, 2013.

[19] A. J. Lotka, *Elements of Physical Biology*, Williams and Wilkins, 1925.

[20] V. Volterra, "Variations and fluctuations of the number of individuals in animal species living together," *Journal du Conseil International pour L'Exploration de la Mer*, vol. 3, no. 1, pp. 3–51, 1928.

[21] M. Li, P. N. Tran, D. Wang, and A. Timm-Giel, "Radio resource allocation in LTE using utility functions based on moving average rates," in *Proceedings of the 2014 IEEE Wireless Communications and Networking Conference, WCNC 2014*, pp. 1891–1896, Turkey, April 2014.

[22] W. Gong, X. Wang, M. Li, and Z. Huang, "Round-robin resource sharing algorithm for device-to-device multicast communications underlying single frequency networks," in *Proceedings of the 21st International Conference on Telecommunications, ICT 2014*, pp. 191–195, Portugal, May 2014.

[23] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wang, and T. Q. S. Quek, "Resource allocation for cognitive small cell networks: a cooperative bargaining game theoretic approach," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3481–3493, 2015.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 4860212, 8 pages
https://doi.org/10.1155/2018/4860212

Research Article

Fronthaul for Cloud-RAN Enabling Network Slicing in 5G Mobile Networks

Line M. P. Larsen, Michael S. Berger, and Henrik L. Christiansen

Department of Photonics Engineering, Technical University of Denmark, 2800, Denmark

Correspondence should be addressed to Line M. P. Larsen; lmph@fotonik.dtu.dk

Received 20 April 2018; Revised 2 July 2018; Accepted 17 July 2018; Published 28 August 2018

Academic Editor: Pei Xiao

This work considers how network slicing can use the network architecture Cloud-Radio Access Network (C-RAN) as an enabler for the required prerequisite network virtualization. Specifically this work looks at a segment of the C-RAN architecture called the fronthaul network. The fronthaul network required for network slicing needs to be able to dynamically assign capacity where it is needed. Deploying a fronthaul network faces a trade-off between fronthaul bitrate, flexibility, and complexity of the local equipment close to the user. This work relates the challenges currently faced in C-RAN research to the network requirements in network slicing. It also shows how using a packet-switched fronthaul for network slicing will bring great advantages and enable the use of different functional splits, while the price to pay is a minor decrease in fronthaul length due to latency constraints.

1. Introduction

Emerging technologies paving the way towards the next generation, 5G, mobile networks include the very promising concept of network slicing. Network slicing describes how one physical network is divided into multiple logical networks, referred to as network slices. One network slice can have specific capabilities related to one specific service, like one slice for Internet of Things (IoT), as this service has specific requirements to the network. Another benefit of network slicing is that it is possible for several operators to share the same physical network and thereby save cost for deployment and maintenance of the physical network equipment. Different network slices have different requirements to the network, because different applications have different requirements to the network [1]. The 3rd Generation Partnership Project (3GPP) describes in [2] how network slicing is envisioned to provide different capabilities on different slices. These capabilities can be related to the three primary 5G drivers [3]:

(i) Extreme Mobile Broadband (eMBB) will support use cases like "shopping mall" requiring 300 Mbps experienced Downlink (DL) throughput and at least 95% availability and reliability for all applications [3].

(ii) Massive Machine-Type Communication (mMTC) will support use cases like "massive amount of geographically spread devices" requiring up to 1,000,000 devices per km^2 and 10 years of battery life [3].

(iii) Ultrareliable Machine-Type Communication (uMTC) will support use cases like "autonomous vehicle control" requiring latency below 5 ms and 99.999% availability and reliability [3].

These examples show how different the requirements can be, to different slices within network slicing. The same physical resources need to be able to carry very different demands, which do not only require complex QoS management but also put huge requirements to the physical network that should be able to handle it. Network slicing requires a virtualization of the network to be able to run several logical networks on top of the physical network [4].

Cloud-Radio Access Network (C-RAN) is a promising network architecture which can be used to enable virtualized networks and network slicing. In C-RAN the base station functions known from the protocol stack are divided into a Distributed Unit (DU) and a Centralized Unit (CU). The DU is located close to the antenna in the antenna mast and is thereby close to the user, where the CU can be located in a datacenter benefitting from high processing powers.

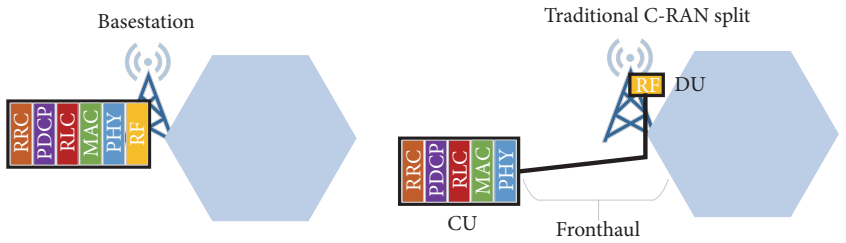


Figure 1: Function allocation in a base station and in the traditional C-RAN split.

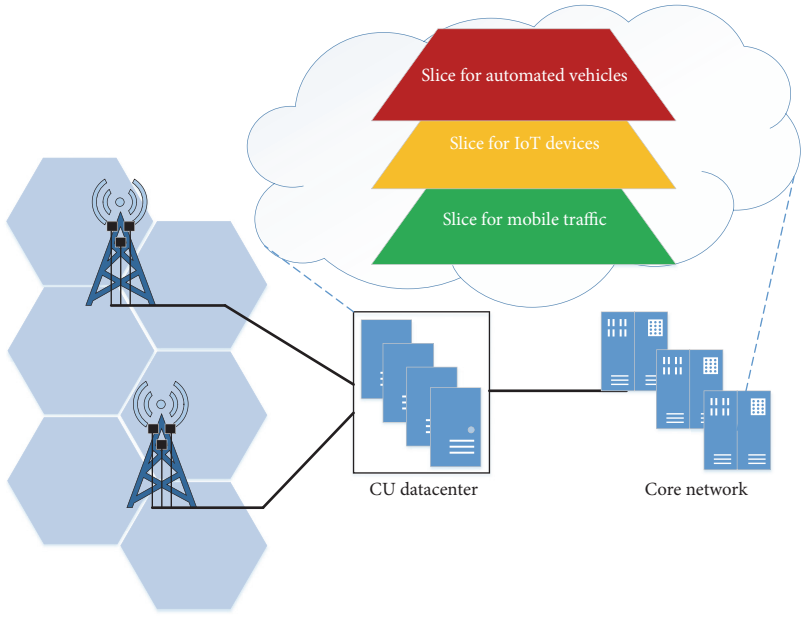


Figure 2: Network slicing in Cloud-RAN.

The exact location of the separation between these two entities is referred to as the functional split. The DU and the CU are connected using a so-called fronthaul network. The simplest division between DU and CU leaves only the Radio Frequency (RF) functions in the DU. This division will be referred to as the traditional split throughout this paper and the division of functions in the traditional split is illustrated in Figure 1.

Figure 1 illustrates the traditional split where the base station functions are illustrated using the LTE protocol stack, as no protocol stack exists for 5G at the time of writing. In the traditional split, raw in-phase quadrature (IQ) samples are transported on the fronthaul link, resulting in a very high and constant bitrate. Using the traditional split, the IQ samples need a special protocol for transport over the fronthaul network. Several options exist including the widely used Common Public Radio Interface (CPRI) [6]. Originally, CPRI was intended for point-to-point transmission, but this makes the fronthaul link very inflexible, as each CU/DU pair requires its own fiber connection. Seen from a network slicing perspective, this solution is not very flexible, as each slice gets a static amount of capacity assigned for fronthaul transmission.

CUs from several sites can be centralized in the same datacenter which is an enabler for modern network virtualization techniques. This way, processing functions are gathered in one place, the CU-datacenter, which can be virtualized. Network functions virtualization moves the network processes into software, and, instead of the functions running at a base station, they will be able to run at any server [7]. Virtualization of several functions is an important enabler for network slicing. The situation is illustrated in Figure 2.

Figure 2 illustrates how network virtualization of the CU-datacenter is used to run several logical network slices on top of one physical network. The logical slices serve different purposes as each of them complies with different network requirements. Virtualizing C-RAN brings benefits in scalable management of processing resources and enables network programmability [8]. This work looks into how network slicing can be enabled using C-RAN with a special focus on the issues raised in future fronthaul networks. The remainder of this paper contains an overview of the current trends being investigated within fronthaul deployment, a comprehensive description of the fronthaul challenges that are still under research, and a discussion of the options existing for C-RAN deployment.

Table 1: Comparison of a circuit switched and a packet switched fronthaul link.

	Circuit switched fronthaul	Packet switched fronthaul
Pros	Guaranteed bitrate.	Adaptable to non-uniform traffic.
Cons	Load independent of cell load.	Delay can occur
Capacity	Guaranteed	Depends on technology
Timing	Guaranteed	Depends on technology
Synchronization	Guaranteed	Depends on technology
QoS	Dedicated user channel	Shared transmission
Service Guarantee	Dedicated resources	Delay can occur
Multiplexing	TDM or WDM	Statistical multiplexing
Resource utilization	Constant use of resources	Improved

2. Recent Trends in Mobile Fronthaul

The mobile fronthaul network has been considered in several papers, investigating different solutions and options. C-RAN and network slicing are two concepts that have already been combined in several works. Reference [5] contributes with a dynamic network slicing scheme for multitenant C-RANs. In [9] a demo of network slicing using C-RAN is introduced which aims at efficiently sharing the bandwidth resources among different slices.

A large survey on cloud-based services in [8] states the benefits of virtualized networks and argues for the use of the C-RAN architecture due to its scalability and high system capacity. The standardization of C-RAN and the division of functions between the CU and DU are a currently on-going process, where 3GPP contributes with [10] and IEEE have established the 1914 Next Generation Fronthaul Interface (NGFI) working group [11]. Also different industry alliances are looking into the subject including the CPRI consortium [12], NGMN alliance [13], and Small Cell Forum [14]. In [15], the authors argue for using the Ethernet network as the new fronthaul, as it is already deployed and brings several benefits. Reference [16] reports on the performance of different functional splits over a bridged Ethernet network; results show fronthaul processing delays and how these are affected by different types of traffic flows. Different papers and organizations use different naming or numbering of the functional splits. To illustrate the functional splits from a more generic point of view, this paper will not refer to any names, but only to the locations of the functional splits in the LTE protocol stack. And only those functional split locations with specific fronthaul abilities will be mentioned, as many more exist. For a complete overview of the options currently researched are referred to in [17].

3. Fronthaul Link Type

The performance of a fronthaul network depends on whether a circuit-switched solution or a packet-switched solution is chosen. The capabilities of a circuit- and a packet-switched fronthaul network are summarized in Table 1.

Circuit-switched solutions, for example, Optical Transport Network (OTN), provide a constant use of resources, as the connections are always using the same amount of capacity, which is statically assigned. This means that queuing will be very limited as the resources are always reserved and the network provides a stable connection in terms of capacity, timing, and synchronization. OTN is a circuit-switched solution that provides protection and multiplexes several transmissions using WDM [18]. The pros of using OTN are that it uses Forward Error Correction (FEC) and it can measure the Bit Error Rate (BER). But when using OTN, or another circuit-switched technology in the fronthaul network, the capacity is fixed to the already assigned carriers and is not able to assign extra capacity to establish or scale slices dynamically.

In packet-switched technologies, it is possible to dynamically assign capacity when needed. Using statistical multiplexing, a packet-switched network is adaptable for a variable load on the fronthaul, because multiple connections are multiplexed into one fiber. Ethernet is an example of a packet-switched fronthaul technology. Ethernet allows sharing of the network infrastructure through standardized virtualization techniques and, through its packet-switched operation, it will save resources on the fronthaul link using statistical multiplexing gain [15]. This makes Ethernet able to flexibly allocate resources and, when considering network slicing, dynamically reserve resources for specific slices. Ethernet has several advantages to be used for fronthaul; it is flexible, cost-effective, and widely used. The Ethernet network has one problem though, before it can be used as the new fronthaul network; in a packet-switched technology, traffic is more eager to queue. Queuing traffic can create unwanted delay and jitter in a mobile network, and therefore the industry is looking into solutions to avoid that using for example carrier grade management. In Ethernet the latency will also be affected by the number of switches to be passed through; this is the reason why Time Sensitive Networking (TSN) is providing options for traffic prioritization in the standard for frame preemption [19].

The trend is pointing towards using Ethernet as the fronthaul network [15], and this will be the focus in the remainder of this paper.

4. Challenges in Mobile Fronthaul

The introduction of network slicing requires a virtualization of the network and an efficient use of resources in the

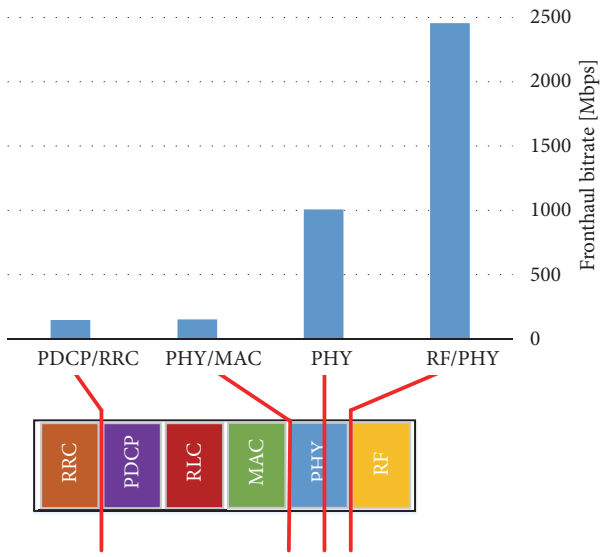


Figure 3: Fronthaul bitrates for selected functional splits. Note that, as per formulas in [5], the CPRI linecode is included in the bitrate for "RF/PHY" split, but not in the "PHY" split.

fronthaul network. This chapter looks into how Ethernet can be the solution to that.

4.1. Functional Split. A packet-switched technology, like Ethernet, does not obtain many benefits when transmitting data requiring a constant bitrate. Therefore, with the user bitrates highly increasing in 5G [2], the traditional split is not a sustainable solution due to the very high and constant fronthaul bitrate. The term functional split defines different options for splitting up the functions in the protocol stack. The trend points towards including more functions in the DU compared to the traditional split. When only a few functions are left in the DU, the signal is raw and the fronthaul bitrate is high with a constant load. Adding more functions in the CU will decrease the fronthaul bitrate and increase the fronthaul flexibility, as the signal gets more processed before the transmission, resulting in a fronthaul bitrate that will vary with the user load. But the low and variable bitrate comes with a cost; the DUs become more complex and thereby more difficult to install and maintain. A variable bitrate on the fronthaul network is crucial for Ethernet to perform dynamical resource allocation, and the lower the bitrate is, the more the resources can be multiplexed into the same fiber. The fronthaul bitrate can be calculated by looking at what type of data is actually being transmitted on the fronthaul link, which is different depending on what functional split is chosen. To state an example of the huge difference between the different functional splits, the fronthaul bitrates for different functional split options are illustrated in Figure 3. The fronthaul bitrates are calculated based on formulas from [14] and considering a 20 MHz LTE carrier using two antenna ports and 64 QAM modulation. The fronthaul bitrates are calculated for four different splits: the traditional split between the RF and physical layer (RF/PHY), a split in the physical layer (PHY), a split between the physical layer and the Media Access Control (PHY/MAC), and a split between the Packet Data Convergence Protocol and Radio Resource Control (PDCP/RRC). The bitrates and the corresponding split location in the LTE protocol stack. The splits and their corresponding fronthaul bitrates are illustrated in Figure 3. Here functions on the left side of the red line are included in the CU and functions on the right side are included in the DU.

Figure 3 illustrates the locations of four different functional split options, and their corresponding fronthaul bitrates. The bitrate for the RF/PHY option is very high, and the bitrates for the PDCP/RRC and PHY/MAC options are very low. The functional split will also determine whether the bitrate on the fronthaul link is constant or varies with user load. This is determined by the amount of functions left in the DU, i.e., the amount of signal processing from the physical layer taking place in the DU. Looking at Figure 3, then the RF/PHY split has a constant load, where the PHY/MAC, PDCP/RRC splits have bitrates varying with user load on the fronthaul link. Whether the fronthaul bitrate is variable or not for the PHY split is depending on where in the physical layer the PHY split is located. Figure 4 illustrates the functional blocks within the physical layer and what types of data is transported in-between them. The blocks that affect the fronthaul bitrate the most are highlighted and will be further discussed; the remaining blocks are included in the figure for completeness. Read from the right side, data is received/transmitted from the RF block represented by IQ symbols. The red line illustrates how functional splits that includes the RE (de)mapper in the DU have a variable load on the fronthaul link, where functional splits located closer to the RF block have a constant load on the fronthaul link.

Read from the right towards left, Figure 4 illustrates how the DL signal received by the antennas are transmitted via the antenna ports into processing in the physical layer before it is further sent to the MAC using transport blocks, and reverse for the receiver. In the RF block, the signal is up/down converted and sampled. When entering the PHY block, the cyclic prefix is removed and sent into the Fast Fourier Transformation (FFT) where the signal is converted into subcarriers in the frequency domain. The FFT process induces a large reduction on the fronthaul bitrate, which in LTE corresponds to 40%, due to removal of guard subcarriers [20]. The Resource Element (RE) mapper maps between subcarriers and symbols. In this process, the unused subcarriers forwarded from the FFT can be detected. This way the RE mapper makes the signal vary with the user load [20]. Therefore, DUs that include the RE mapper will have a variable bitrate on the fronthaul link. And DUs that do not include the RE mapper will have a constant bitrate on the fronthaul link; i.e., functional splits on the left side of the red line have variable bitrate, and functional splits on the right side have constant bitrate on the fronthaul link. Another process in PHY that will further reduce the fronthaul bitrate is the modulation. When modulating the signal, a certain amount of bits will be mapped into one symbol, depending on the order of modulation [21]. The order of modulation will then determine how much the signal is reduced.

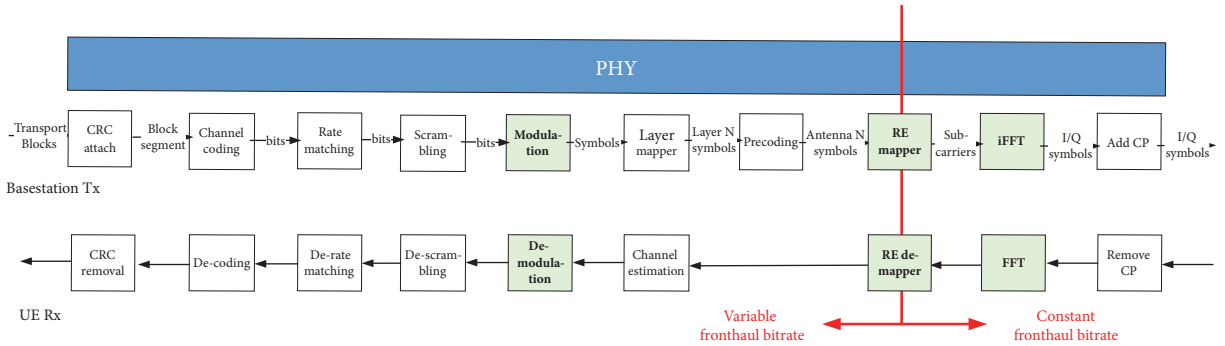


Figure 4: Functional blocks in the physical layer, shown for both uplink and downlink direction.

Using transport blocks the signal is transmitted from the physical layer into the MAC. In the LTE MAC, the Hybrid Automatic Repeat Request (HARQ) process is located together with a scheduler determining the cooperation among DUs [10]. The HARQ process is very time critical [21] and this puts large latency constraints on the fronthaul network when this process is located in the CU-datacenter. The scheduler on the other hand is beneficial to have in the centralized CU-datacenter for better management of the DU resources. Some functional splits are proposed to separate the scheduler into a local scheduler block in the DUs and a central scheduler in the CU-datacenter [10].

4.2. Fronthaul Delay. It is crucial for the fronthaul link to comply with different delay requirements, both because different functions within the protocol stack have different delay requirements and because different applications have different requirements to the response time. In network slicing, where all services rely on one network, the network needs to be compatible with the worst case situations. Using a packet-switched network, an extra delay can be expected due to queuing or a slight delay added in frame preemption for the high priority packets [19]. The delay on the fronthaul link is crucial to determine the length of the fronthaul network, as the distance between a DU and the CU-datacenter is determined by the maximum delay. Since certain functions and services determine the delay limits in the network, this delay will also determine the maximum distance between the DU and the CU-datacenter. The delay is given in

$$Delay = 2 \cdot transmission\ delay + processing\ delay \quad (1)$$

The transmission delay is calculated as

$$Transmission\ delay = \frac{Packet\ size}{fronthaul\ bitrate} \quad (2)$$

The maximum Ethernet packet size is 1500 bytes. The fronthaul bitrates are calculated based on formulas in [14].

The Round Trip Time (RTT) describes the time from a request is sent until a message is received, and therefore it also includes the propagation delay, i.e., the time for the request to propagate through the given medium with a certain length. The RTT is given in

$$RTT = delay + 2 \cdot propagation\ delay \quad (3)$$

The RTT delay needs to be compliant with the delay requirements for the specific service that it is running, and when an Ethernet fronthaul is assumed, the delay for queuing and processing through Ethernet needs to be considered. Therefore a delay for one Ethernet switch is added as Dsw and multiplied by the number of switches:

$$RTT > delay + 2 \cdot Dsw \cdot \#\ switches + 2 \cdot propagation\ delay \quad (4)$$

The delay in one switch, Dsw is depending on the type of switch and the packet size. The following equation is an assumption for Dsw assuming a Gb Ethernet switch is used:

$$Dsw = \frac{Packet\ size}{1 \cdot 10^9} \cdot 1000\,ms \cdot switch\ processing\ time\ and\ queuing\ delay \quad (5)$$

The total switch processing time and queuing delay can be adjusted according to amount of traffic and priority packets. In this work, it is assumed to be factor 3.

The propagation delay is calculated as

$$Propagation\ delay = propagation\ delay\ per\ km \cdot fronthaul\ range \quad (6)$$

The distance between the DU and CU, the fronthaul range, can be determined by

$$Fronthaul\ range \leq \frac{(RTT - delay)}{propagation\ delay\ per\ km} \quad (7)$$

As an example, eMBB is considered as the service giving the delay requirements, here the maximum latency for the user plane is 4 ms [22], corresponding to 8 ms RTT. The fiber propagation delay is 10 μs/km [23].

Figure 5 illustrates how much the fronthaul range is affected when the transmitted data has to pass through different numbers of switches. Depending on the processing delay, the fronthaul range can be several hundred of km. It must be considered that the processing time might be lower in the CU-datacenter compared to the standalone DU. Therefore the processing delay has to be distinguished

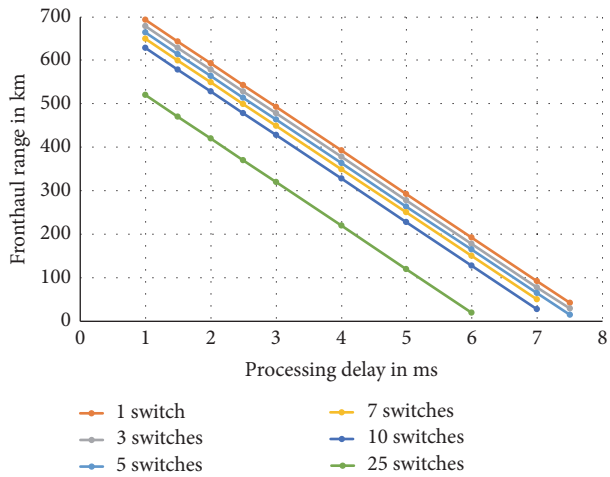


Figure 5: Fronthaul range corresponding to different amounts of fronthaul delay illustrated for different numbers of switches in the fronthaul network.

between the amount of functions being processed in the DU and in the CU-datacenter. The amount of switches the data needs to pass on the way between the DU and the CU-datacenter will most likely depend on the population density in the current area, as a higher population density might incur more switches in the area. Looking at Figure 5, then the difference between passing one switch and 25 switches gives a difference in the fronthaul range of approximately 170 km.

Some functions within the LTE protocol stack have higher requirements to the max delay; this is for example the HARQ process located in the MAC, which is limited by a RTT of 5 ms [21]. And because LTE might need to coexist with other RATs on its own network slice, HARQ is also considered here. In splits where the HARQ process is located in the CU-datacenter, the signal needs to be transferred over the fronthaul network and back within the 5 ms RTT. The RTT of the HARQ located in the DU and the CU, respectively, is illustrated in Figure 6.

Figure 6 shows how the distance to overcome when the HARQ process is located in either the CU or the DU. When the HARQ process is located in the DU, the distance to the user is very short and the signal only has to travel from the user to the DU and back within the 5 ms of RTT. Therefore, the latency for the splits where the HARQ process is included in the DU is more relaxed and results in a much longer fronthaul range. The location of the MAC and HARQ process in the LTE protocol stack is illustrated in Figure 7.

Due to the HARQ process's delay limitations, three delay scenarios exist: one for the splits before the RE mapper, having a constant load on the fronthaul link, one for the remaining splits where the HARQ is located in the CU-datacenter, and one for the splits where the HARQ is located in the DU and the latency is limited by a certain service. The latter one has already been described.

The delay budget for the splits before the RE mapper, with no Ethernet delay added as the connection is expected to be circuit-switched:

$$5\,ms > delay + 2 \cdot propagation\ delay \tag{8}$$

The delay budget for the split options after the RE mapper, where the HARQ is still located in the CU, here an Ethernet fronthaul is assumed and therefore a delay for one Ethernet switch is added as Dsw and multiplied by the number of switches:

$$5\,ms > delay + 2 \cdot Dsw \cdot \#\ switches + 2 \cdot propagation\ delay \tag{9}$$

Figure 8 illustrates the fronthaul range limited by the HARQ process. The number of switches is 5.

Figure 8 shows how the HARQ requirements affect the range of the fronthaul link. The figure also shows how different processing delays will affect the fronthaul range. Applications that have larger requirements to the delay need to follow this limitation if they use the traditional functional split or a split where the HARQ is located in the CU. Applications that have the HARQ located in the DU have larger requirements to the delay resulting in a much longer fronthaul range illustrated in Figure 5. Figure 8 clearly shows that the delay added by the Ethernet network affects the fronthaul length, as the options "Before RE map" and "HARQ in CU" have a clearly division. Theoretically, the processing delay should be lower for the functions implemented in the CU-datacenter, as those have higher processing powers available compared to the single DUs at the cell sites. This would in that case affect the RTT, but this has not been taken into consideration in these calculations. Figure 8 shows how using an Ethernet network as fronthaul has a minor impact on the delay and thereby the fronthaul range.

5. Discussion

In a physical network used for network slicing, one network has to carry the traffic from several logical networks, each having very different requirements to the physical network. The one physical network needs to be able to run all extreme scenarios at the same time. It must have high capacity in terms of bitrates and number of supported devices and it must be extremely resilient and support ultralow latency, all at the same time for the network resources to be utilized on the right slices. C-RAN networks open up for the opportunity to share processing resources and allocate extra resources where they are needed as several functions are incorporated in a datacenter. But when using a circuit-switched fronthaul, resources are statically assigned and provide a dedicated transmission. Therefore a trade-off exists between latency added in packet queues and dynamical resource allocation in the fronthaul network. The trend points towards using the already established Ethernet network as fronthaul. If a solution using Ethernet or another packet-switched fronthaul technology is chosen, the capabilities provided by TSN will be highly necessary to prioritize the functions with very strict latency requirements. In Ethernet the physical resources can be utilized in a more efficient manner by flexibly allocating capacity within the fibers and the switches corresponding to a specific slice's demands. Using a flexible resource allocation, it is optional whether TSN is used in one slice or not. It is also optional what functional split is used in a specific slice

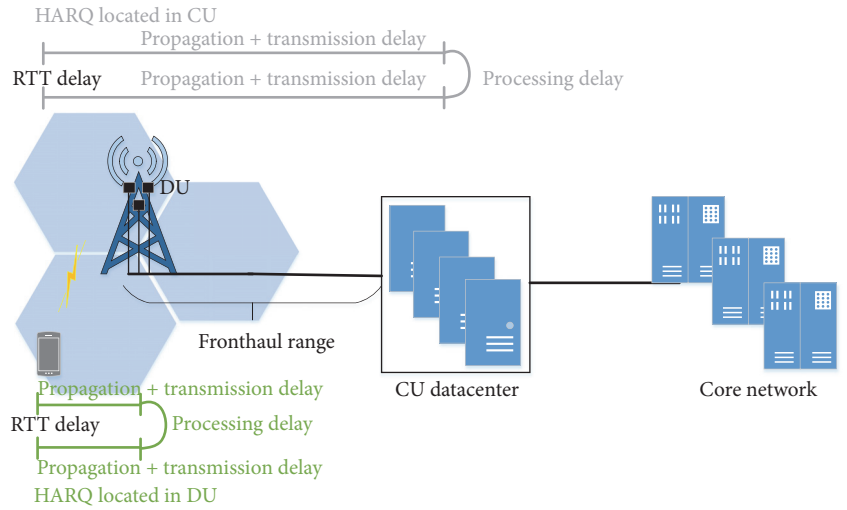


Figure 6: RTT response when HARQ is located in the CU-datacenter and when HARQ is located in the DU.

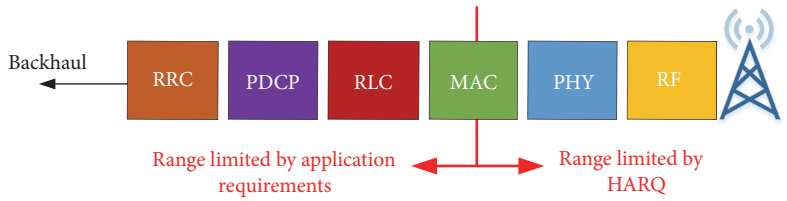


Figure 7: LTE protocol stack illustrating range limitations for splits before and after MAC/HARQ process.

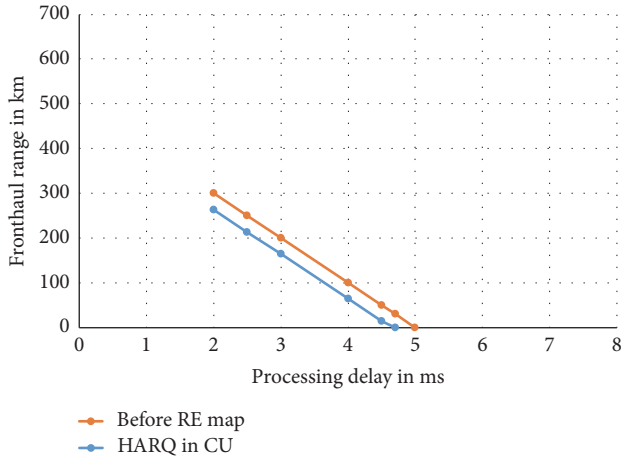


Figure 8: Fronthaul range for LTE scenarios complying with HARQ requirements.

and so on. One big downside of using a packet-switched network is usually that the delay is enhanced, but due to the calculations prior to Figure 8, the delay is minor. The length of the fronthaul, referring to Figures 5 and 8, is depending on the delay requirements and the available processing powers.

A crucial parameter is to decide which functional split to use, or in which situations a certain functional split shall be used. Different functional splits have different pros and cons, which is partly related to what benefits the different functions have when they are local, close to the user or when they are centralized benefitting from large processing powers in the CU-datacenter. Hence different functional splits can be used in different scenarios. Some scenarios might require a large amount of centralized processing and a simple DU, where other scenarios obtain benefits in separating the user plane and control plane. Other functional splits are in-between these extremities: they want the benefits from a simplified DU but they also want a low fronthaul bitrate. Seen from a network slicing perspective it is recommended to use functional splits that have a variable fronthaul bitrate and runs over a packet-switched network. Considering the primary 5G drivers from [3], then eMBB will benefit from a variable bitrate on the fronthaul link; as enormous amounts of data needs to be transmitted at peak times, it will also benefit from a large amount of centralized processing, for a more efficient allocation of resources. mMTC has relaxed bitrate requirements; therefore it will benefit from a simple DU for easy deployment and maintenance. uMTC has very strict requirements to the latency, and thereby it requires a network with good traffic management and flexible resource allocation in the fronthaul network for faster transmission, and it will benefit from a centralized MAC scheduler. If C-RAN carries the network architecture for the physical network in network slicing, this one network needs to be compatible with all the requirements in all slices. In this manner it would make most sense to use different functional splits for different slices, in order to obtain the best resource utilization and performance.

6. Conclusion

C-RAN is a promising network architecture enabling virtualization techniques to be used for example for network slicing. In C-RAN the sites are connected, using a fronthaul network, to high capacity datacenters running more or less base station processing functions. Deploying a fronthaul network faces a trade-off between fronthaul bitrate, flexibility, and complexity of the local equipment close to the user. Network slicing introduces the concept of one physical network running several logical networks with different requirements on top. As the different slices can have very large and differentiated requirements to the network, the physical network needs to be equipped to handle extreme scenarios. This work showed how using a packet-switched fronthaul, for network slicing will bring great advantages and enable the use of different functional splits, while the price to pay is a relatively minor delay.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the Eurostars "Fronthaul for C-RAN" Project funded by Innovation Fund Denmark.

References

[1] S. Sharma, R. Miller, and A. Francini, "A cloud-native approach to 5G network slicing," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 120–127, 2017.

[2] 3GPP TS 22.261 V16.3.0, "Service requirements for the 5G system; Stage 1," 2018.

[3] A. Osseiran, J. F. Monserrat, and Y. P. Marsch, *5G Mobile and Wireless Communications Technology*, Cambridge University Press, Cambridge, UK, 2016.

[4] T. Yoo, "Network slicing architecture for 5G network," in *Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC '16)*, pp. 1010–1014, IEEE, Jeju Island, Korea, October 2016.

[5] Y. L. Lee, J. Loo, T. C. Chuah, and L. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2146–2161, 2018.

[6] CPRI Consortium, "CPRI Specification V7.0 Common Public Radio Interface (CPRI); Interface Specification," 2015.

[7] B. Chatras, U. S. Tsang Kwong, and N. Bihannic, "NFV enabling network slicing for 5G," in *Proceedings of the 20th Conference on Innovations in Clouds, Internet and Networks (ICIN '17)*, pp. 219–225, March 2017.

[8] G. P. Xavier and B. Kantarci, "A survey on the communication and network enablers for cloud-based services: state of the art, challenges, and opportunities," *Annals of Telecommunications-Annales des Télécommunications*, vol. 73, no. 3-4, pp. 169–192, 2018.

[9] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar, "DEMO: SDN-based network slicing in C-RAN," in *Proceedings of the 15th IEEE Annual Consumer Communications & Networking Conference (CCNC '18)*, pp. 1-2, 2018.

[10] 3GPP, "3GPP TR 38.801 V14.0.0 (2017-03): Study on new radio access technology: Radio access architecture and interfaces," 3GPP, 2017.

[11] IEEE 1914, http://sites.ieee.org/sagroups-1914/, Accessed 2018-13-02.

[12] "eCPRI Specification V1.1 Common Public Radio Interface: eCPRI Interface Specification".

[13] NGMN, "NGMN Overview on 5G RAN Functional Decomposition," 2018.

[14] Small Cell Forum, "Solving the HetNet puzzle Small cell virtualization functional splits and use cases".

[15] N. J. Gomes, P. Sehier, H. Thomas et al., "Boosting 5G through ethernet: how evolved fronthaul can take next-generation mobile to the next level," *IEEE Vehicular Technology Magazine*, vol. 13, no. 1, pp. 74–84, 2018.

[16] P. Assimakopulos, G. S. Birring, M. K. Al-Hares, and N. J. Gomes, "Ethernet-based fronthauling for cloud-radio access networks," in *Proceedings of the 19th International Conference on Transparent Optical Networks (ICTON '17)*, pp. 1–4, July 2017.

[17] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Communications Surveys & Tutorials*, 2018.

[18] Y. Ma, X. Huo, J. Li, Xiaomu, and Jingwen, "Optical solutions for fronthaul application (invited)," in *Proceedings of the 14th International Conference on Optical Communications and Networks (ICOCN '15)*, pp. 1–3, July 2015.

[19] IEEE computer Society, "IEEE Std 802.1Qbu: Frame Preemption," 2016.

[20] D. Wübben, P. Rost, J. S. Bartelt et al., "Benefits and impact of cloud computing on 5 G signal processing: flexible centralization through cloud-RAN," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, 2014.

[21] M. Sauter, *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*, 2011.

[22] 3GPP, "TR 38.913 V14.3.0 (2017-06): Study on Scenarios and Requirements for Next Generation Access Technologies".

[23] D. Chitimalla, K. Kondepu, L. Valcarenghi, M. Tornatore, and B. Mukherjee, "5G fronthaul-latency and jitter studies of CPRI over ethernet," *Journal of Optical Communications and Networking*, vol. 9, no. 2, pp. 172–182, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 7870164, 17 pages
https://doi.org/10.1155/2018/7870164

Research Article

Modelling of Multiservice Networks with Separated Resources and Overflow of Adaptive Traffic

Mariusz Głąbowski, Damian Kmiecik, and Maciej Stasiak

Poznań University of Technology, Polanka 3, 60-965 Poznań, Poland

Correspondence should be addressed to Mariusz Głąbowski; mariusz.glabowski@put.poznan.pl

Received 30 March 2018; Revised 29 June 2018; Accepted 10 July 2018; Published 14 August 2018

Academic Editor: Dejan Vukobratovic

The article proposes a new method of determining traffic characteristics of multiservice overflow systems that carry adaptive traffic. When the total offered load in primary resources exceeds a certain value, this type of traffic is admitted for service with lower bitrate. A particular attention is given in the article to a method for a determination of the parameters of traffic that overflows to secondary resources as well as to the way adaptive traffic is serviced. The method takes into consideration three possible types of traffic: Erlang, Engset, and Pascal traffic. It is based on a generalization of Hayward's concept and its application to model systems with adaptive traffic with threshold compression. The method can be used for optimal dimensioning of logical networks (slices) in modern mobile systems due to possibility of analytical determination of grade of service parameters (blocking probability, carried traffic, and network load). To verify the accuracy of the proposed model the results of analytical calculations, obtained on the basis of the proposed model, are then compared with the results of simulation experiments for a number of selected structures of overflow systems that service adaptive traffic. The results of the study demonstrate high accuracy of the proposed theoretical model.

1. Introduction

The mechanism of traffic overflow has been used in telecommunications networks for over seventy years as one of the oldest optimization techniques for traffic distribution. The overflow mechanism is initiated when the resources of a given system (the so-called primary resources (PR)) are occupied and, in consequence, unavailable for new calls. These calls, however, can be admitted for service by other systems that have free resources (the so-called secondary resources (SR)) available to handle new calls [1]. In the 1950s, the introduction of cross-connect switching systems to telecommunications networks allowed overflow mechanisms to be applied in network engineering. These mechanisms for single-service networks have been thoroughly described and analyzed in such classic works as [2–12]. In [2, 3], the Equivalent Random Traffic (ERT) method is developed for a determination of the blocking probability in single-service overflow systems. This method takes advantage of the first two moments of traffic that overflows from PR to SR, i.e., the mean value and variance. This, in turn, provides a basis for the so-called equivalent resources to be defined. Equivalent resources allow then the blocking probability in an overflow system to be directly determined. An interesting method for a description of single-service overflow systems is proposed in [10]. The method is based on a modification of Erlang's formula, in which the capacity of secondary resources and the value of traffic offered to the overflow group are divided by the peakedness factor of overflow traffic, i.e., by the ratio of the variance of overflow traffic to its mean value. Such an approach has significantly simplified the way the blocking probability in overflow systems can be determined as compared to the ERT method. The works [5, 13, 14] point out the possibility of an approximation of the call stream that overflows from the primary resources in a call stream of Pascal type. In [15, 16] systems with mutual traffic overflow are considered and discussed.

A commercial application of the first multiservice Integrated Services Digital Network (ISDN) in continuously carried out until now within the context of new technologies and network standards that are being successively introduced. The works [17–19] propose a description of overflow traffic

with Markov-Modulated Poisson Process (MMPP processes), whereas in [20, 21] the Batched Poisson Process (BPP) is used. The authors of [20, 22, 23] consider a possibility of an application of traffic stream with an assumed value of the peakedness factor. The papers [24–26] propose a number of engineering methods for dimensioning multiservice systems with overflow traffic. These methods are based on an application of the approach [10] with regard to multiservice full-availability resources (FAR full availability resources), the so-called full-availability groups (FAG) [1, 27]. The full-availability approach means that a new call will be admitted for service if the system has sufficient capacity (resources) necessary for this call to be serviced. FAR with offered mixture of Erlang, Engset, and Pascal traffic streams with varied bitrates can be described on the basis of multiservice Markov processes that, when solved, provide simple recurrence formulas [28, 29] (for a mixture of Erlang traffic) and [30, 31] (for a mixture of Erlang, Engset and Pascal traffic). Dividing the values of traffic of individual classes offered to the secondary resources by the corresponding peakedness factors makes it possible to model multiservice SR and, in consequence, the blocking probability of an overflow system. In [32], to model overflow systems a two-dimensional convolution model is proposed, while in [33] the overflow system is approximated by ideal nonfull availability resources (called Erlang Ideal Grading (EIG) in the literature of the subject [34]). In models [35–37] overflow systems are considered in which overflow traffic changes the service parameters in the secondary resources, such as the service time and bitrate. In the study, a change in the parameters is not, however, related to network mechanisms for traffic shaping.

Present-day multiservice networks, including 4G and 5G mobile networks, are packet networks (based on IP (Internet protocol) at Internet Layer and TCP (Transmission Control Protocol) at transportation layer in which packet streams undergo different traffic shaping mechanisms. One of the most widely used mechanisms used in TCP/IP networks, both wired and mobile, are the mechanisms of threshold and thresholdless traffic compression. Traffic that is subject to the operation of these mechanisms is called elastic [38] (based on TCP) or adaptive traffic [39] (based on RTP/UDP (Real Time Protocol/User Datagram Protocol)), respectively. Compression mechanisms lead to a decrease in the bitrate of new or currently serviced packet streams and, in consequence, make it possible for a larger number of streams to be serviced. A decrease in the bitrate can be accompanied with an extension of service time, which occurs, for example, in the case of an execution of a service in which all data has to be transferred in its entirety (elastic traffic). In other cases, the service time is fixed (adaptive traffic). Elastic and adaptive traffic can be executed by threshold and thresholdless compression mechanisms. Thresholdless compression mechanisms influence calls that are being serviced, whereas threshold compression mechanisms reduce bitrate of streams in the admission stage of new calls with the help of the Call Admission Control (CAC) function. When this service has been initiated, the bitrate determined by the CAC function is not changed any more. In practice, elastic traffic is executed in a thresholdless manner, since it is characteristic for all services that employ the TCP protocol. These are typically nonreal time services, in which all data has to be transmitted. Hence, a decrease in the bitrate will be accompanied by an extension of the service time. Adaptive traffic, in turn, is typically executed in the threshold manner and is characteristic for real time services that employ the UDP protocol. Since the UDP protocol itself does not have an ability of influencing a packet source, it was necessary to introduce the RTP (Real Time Protocol) and RTSP (Real Time Streaming Protocol) protocols in upper layers. The protocols allow rapid changes in the bitrate speed of generated packet streams to be executed without any extension of the service time.

The work [40] proposes a multiservice FAR model that services thresholdless elastic traffic with finite compression. The notion of finite compression means that the allocated bitrate for a given flow (call) can be decreased within certain boundaries. In [41], this model is generalized to include infinite thresholdless compression in which flows always decrease their bitrates as soon as resources are occupied. In such a system, the phenomenon of call blocking will never occur even though the bitrates of serviced calls can, with large loads of the system, tend towards zero. The analytical base for the models [40, 41] is multidimensional reversible Markov processes with strictly specified values for service streams in individual states. They directly determine the distribution of FAR resources between serviced call classes. This distribution, resulting from a Markov process, is compliant with the so-called balanced fairness algorithm [42–44]. The algorithm leads to a state-dependent Markov process service of all traffic classes offered to the system. In [45], the model [40] is expanded to include service of thresholdless elastic and adaptive traffic.

The first FAR model with threshold compression is proposed in [46, 47]. In this model, developed for Erlang elastic traffic, one threshold is used (the so-called Single Threshold System (STS)). When the threshold is crossed by input signal, the bitrates of new calls that arrive are decreased. In [40], STS for elastic Engset traffic is proposed. The work [48] discusses systems with multiple thresholds (the so-called Multi Threshold System (MTS)) and for elastic Erlang traffic, while works [49, 50] discuss systems for Engset traffic and ON/OFF traffic, respectively. The authors of [1, 51] propose MTS for elastic and adaptive Erlang, Engset, and Pascal traffic. In [52], the threshold compression mechanism is replaced with a more complex hysteresis mechanism, in which thresholds are dependent on the direction of changes in the load of the system, i.e., two thresholds are introduced: one for the direction of changes: low loads-high loads; the other for the direction of changes: high loads-low loads. The paper [53] describes a model with double hysteresis and Erlang, Engset, and Pascal traffic.

The problem of determining traffic parameters of multiservice networks and optimal resource allocation in multiservice networks became especially important in the case of 5G mobile networks that rely on virtualization (slicing) of resources [54–57]. One of the fundamental difficulties arises from the necessity of servicing different classes of

traffic streams by a network. The next important element influencing the complexity of the resource management mechanisms in virtual networks (slices) is the fact that resources can be executed both with the utilization of a single physical resource and with the utilization of many physical resources. The initial analysis of the problem related to designing feasible strategies of resource management in multiservice network indicates that the ones of the most effective strategies can be reservation mechanisms of resources, both dynamic (executed online and securing well-balanced access to network resources) and static (executed appropriately ahead of time with time advancement) [1], as well as threshold mechanisms [58, 59] and priority mechanisms [60].

The mechanisms adopted by telecommunications operators and network protocols (e.g., reservation mechanism and compression) require performing appropriate traffic analysis of the operating network systems and their optimal dimensioning. Until recently, the main emphasis has been put on the blocking probability calculation in systems without virtualization. Nowadays, in order to fully determine the influence of resource management mechanisms on the effectiveness of telecommunications networks as well as in order to determine the optimal share of virtual operators in physical resources, it is necessary to work out analytical methods that would enable us to model traffic characteristics of virtual networks with various resource management mechanisms and network protocols. Additionally, in the case of the mobile networks implementing slicing of resources, the very promising technique of optimization of network resources utilization is the overflow technique. The possibility of having dedicated resources (treated as primary resources) for particular traffic streams (flows) allows mobile network operator also to reserve a common resources (treated as secondary resources) for streams that—even after compression—could not be carried by the primary resources.

The first results on modelling networks with traffic overflow and elastic traffic were published in [45, 61]. The model [61] is generalized in [62] to include systems that service thresholdless elastic traffic in which secondary resources have a distributed nature, i.e., are composed of a number of separated resources with identical capacities [63] or differentiated capacities [64]. A model in which both primary resources and secondary resources have distributed nature is proposed in [65].

Until now, no models of overflow systems with threshold traffic compression have been developed, i.e., no model of an overflow system with adaptive traffic in which calls can undergo threshold compression in both primary resources and secondary resources has been developed. This means that, in the case where certain loads of the resources, determined earlier by an adopted set of thresholds, are exceeded, new calls will be admitted for service with decreased bitrates and unchanged service time. The absence of free resources in primary resources results in a situation where a call overflows to secondary resources, whereas lack of free resources in the secondary resources leads to a loss of the call. This article proposes a new model of an overflow system with adaptive Erlang, Engset, and Pascal traffic, executed by threshold mechanisms. The basis for the analytical description of primary and secondary resources will be the resource models proposed in [1, 51]. The accuracy of the proposed model will be verified on the basis of a comparison of the results of the analytical calculations with the results of the simulation experiments carried out for a network with dedicated primary resources and a common secondary resources.

The present article is structured as follows. Section 2 presents a description of the multiservice overflow system for adaptive traffic in which calls undergo the threshold compression mechanism. Section 3 includes a description of primary resources with streaming and adaptive traffic that is compressed in the threshold manner. This section also provides a description of a method for an evaluation of the traffic parameters for traffic that overflows from primary resources. Section 4 provides a method for modeling secondary resources as well as the algorithm for a determination of the blocking probability in an overflow adaptive traffic system. Section 5 provides a comparison of the results of the analytical calculation with the results of the simulation experiments for selected structures of overflow systems. Section 6 sums up the article.

2. Overflow of Adaptive Traffic in Mobile Networks

Literature on modeling networks (systems) with traffic overflow mostly considers systems in which primary resources are composed of a number of independent resources, called primary groups or direct groups, and secondary resources called secondary or alternative groups [66]. The approach adopted for this article employs the following terminology for primary resources (PR (primary resources)) and secondary resources (SR (secondary resources)) [62, 65] that better express the substantial idea of the overflow phenomenon described in the article, i.e., the authors believe that the notions of "virtual resources", "cell resources", or "radio interface resources" will be more understandable by the reader than the notion of "group".

The assumption in the article is that in both primary resources (slices) and secondary resources (slices) certain classes of traffic—related to the RTP/UDP protocols—can independently undergo threshold compression. This means that when given resources exceed the assumed occupancy threshold (load), new calls will be admitted with decreased bitrate that, during the service, will remain unchanged. Since the model aims at a description of adaptive traffic, the call service time is always the same and is independent of the compression process. Absence of appropriate bitrates for calls that are compressed in primary resources will cause, in turn, these calls to overflow to secondary resources. Since the threshold compression mechanism is assigned to given resources, the assumption is that a given flow overflows in noncompressed form and, while in secondary resources, can undergo independent (of primary resources) threshold compression. Absence of free resources in secondary resources leads then to losses in flows (calls) of particular classes.

2.1. Traffic Representation in the System. In modern mobile networks information is transmitted in IP packets. Packets that belong to a given service that is just being executed form streams that can be considered as calls or flows, e.g., [42, 45, 67]. A mathematical analysis of systems in which the internal structure of packet streams is taken into consideration is very complex [68, 69] and frequently leads to solutions that are approximate and of a very limited practical applicability or to simulation solutions [70, 71]. In dimensioning and optimization of network systems, however, the approach based on an analysis at the packet level cannot be applied to an all-encompassing and broad-based network analysis. The analysis at packet level only allows the aggregated parameters of a packet stream to be evaluated, whereas a determination of the characteristics of streams related to a given service in such a mixture of different streams with different characteristics is impossible. Therefore, the packet approach is not workable and inadequate for network operators. In traffic theory of multiservice systems a dominant approach is then the "call level" approach. It is only this approach that makes it possible to develop coherent models for dimensioning and optimization of networks that would make evaluation of the characteristics of individual services with their relevant GoS (Grade of Service), QoS (Quality of Service), and QoE (Quality of Experience) parameters taken into consideration possible. The results of the studies [58, 67] carried out in recent years indicate that streams at the call level can be described or approximated by streams that have "Poisson" nature. The adoption of such an approach makes it possible to discretize the system and to analyze it on the basis of multidimensional Markov processes. Discretization is based on an exchange of the variable bit rate (VBR) of a packet stream that forms a call by a certain constant bit rate (CBR), called equivalent bandwidth (EB) [72, 73]. EB describes then a constant bitrate that is logically allocated to a given call to provide appropriate (accurate) service parameters for this call in the network. Equivalent bandwidth is typically determined heuristically [74, 75], with regard to such parameters as the total capacity of the system, maximum and average bitrate of the packet stream, bit rate variance, maximum packet delay (latency), jitter, and other parameters characteristics for a given network technology [75–78]. Nowadays, in the description of multiservice systems related to modern packet networks (TCP/IP networks in particular), the assumption is that EB of relevant call classes is determined on the basis of the maximum bitrates of packet streams that correspond to calls. Such an approach is compliant with the system dimensioning principle for the highest network load conditions. It should be pointed out at this point, however, that the method for EB determination has no influence on a mathematical model of a system under consideration, while a choice as to the most appropriate method should be subjected to relevant arrangements between the network operator and involved entities or stakeholders that are to execute all the tasks concerning the design, development, and optimization of a network.

The next stage in the discretization process is a determination of the allocation unit (AU) for a given system, also called in literature Basic Bandwidth Unit (BBU) [58, 79, 80]. Most frequently this unit is defined as such a bitrate that a demanded EB of calls of individual classes offered to the system (called demands) is the multiple of the allocation unit. If an assumption is that a demanded EB for calls of individual classes that are offered to the system is determined on the basis of the maximum bitrates of calls of individual classes, then the maximum value of AU can be determined as the Greatest Common Divisor (GCD) of all maximum bitrates of calls offered to the system:

$$\max c_{\mathrm{AU}} = \mathrm{GCD}\left(c_{1,\max}, c_{2,\max}, \ldots, c_{M,\max}\right), \tag{1}$$

where $c_{i,\max}$ is the maximum bitrate of a call of class i, whereas c_{AU} is the bitrate of AU. In the last stage of the discretization process, both the demands of calls of individual classes and the capacity of the system (real or virtual) are expressed in AUs:

$$\begin{aligned} t_i &= \frac{c_{i,\max}}{c_{\mathrm{AU}}}, \\ V &= \left\lfloor \frac{C}{c_{\mathrm{AU}}} \right\rfloor, \end{aligned} \tag{2}$$

where C is the capacity of the system.

Since the value of AU determined on the basis of (1) is the maximum value, then if at a certain stage of considerations noninteger values of demands t_i for calls of particular classes appear (e.g. as a result of threshold traffic compression), then the system can be rescaled by an appropriate decrease in the value of AUs, which will provide integer values for demands of all call classes that are serviced in a system with threshold compression. For example, if the assumption is that as a result of the operation of a certain traffic shaping mechanism in the system, calls of class i with altered bitrate $c_{i,\max,\mathrm{new}}$ will appear, then the value of AUs can be selected on the basis of (1), with new bitrates for calls of individual classes taken into consideration:

$$\begin{aligned} \max c_{\mathrm{AU}} = \mathrm{GCD}\big(&c_{1,\max}, c_{2,\max}, \ldots, c_{M,\max}, c_{1,\max,\mathrm{new}}, \\ &c_{2,\max,\mathrm{new}}, \ldots, c_{M,\max,\mathrm{new}}\big). \end{aligned} \tag{3}$$

In modelling network systems related to TCP/IP networks the frequent assumption is that one AU has the value of 1 bps (or 1 kbps). Such an approach allows modeling of network systems to be greatly simplified [58]. The assumption in this article is that a call of each traffic class is determined by the number of demanded allocation units, calculated earlier on the basis of (2) and (3), that will always be integer numbers. Another assumption in the article is that both primary resources and secondary resources (real or virtual) have a defined capacity, expressed in AUs. Demands of individual call classes, also expressed in AUs, are known.

2.2. Resource Model with Threshold Compression for Adaptive Erlang Traffic. As an appropriate resource model with threshold compression for Erlang traffic, the Multi Threshold System (MTS), discussed in [1] can be used. Full-availability resources (FARs), with the capacity V AUs, are offered m call classes from the set M. For each call of class i ($0 < i \leq$

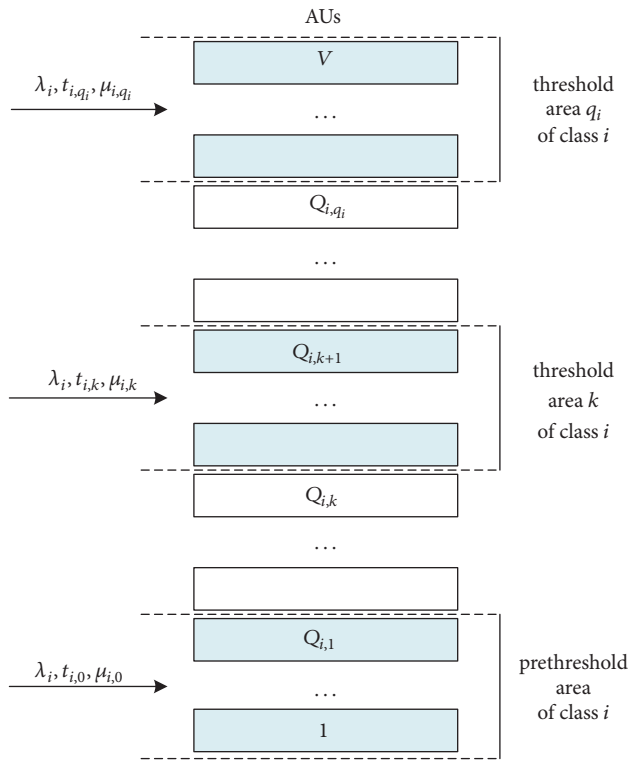


Figure 1: Resource model with threshold compression for adaptive Erlang traffic for class i calls.

m), a set q_i of thresholds $\{Q_{i,1}, Q_{i,2}, \ldots, Q_{i,q_i}\}$ is introduced individually, while $\{Q_{i,1} \leq Q_{i,2} \leq \ldots \leq Q_{i,q_i}\}$. A threshold is understood to be a certain occupancy state in FAR, expressed in the number of occupied AUs. The occupancy area, such that the number of occupied n AUs satisfies the condition $0 \leq n \leq Q_{i,1}$, is called the prethreshold area. If the number of occupied AUs belongs to the postthreshold area k, then the condition $Q_{i,k} < n \leq Q_{i,k+1}$ is fulfilled, where $0 < k \leq q_i$. The postthreshold area q_i satisfies the condition $Q_{i,q_i} < n \leq Q_{i,V}$ (Figure 1). In each postthreshold area k, the number of AUs allocated to service a call of class i is $t_{i,k}$, whereas this parameter in the prethreshold area is $t_{i,0}$. The assumption is that the parameters $t_{i,k}$ satisfy the inequality $t_{i,0} \leq t_{i,1} \leq \ldots \leq t_{i,q_i}$. In the prethreshold area and in each postthreshold area k, offered traffic of class i is described by its own set of parameters $\{\lambda_{i,k}, \mu_{i,k}, t_{i,k}\}$. The parameter $\lambda_{i,k}$ is the intensity of calls of a Poisson stream (Erlang type traffic) and is independent of the FAR occupancy state. Since adaptive traffic is considered in MTS in which a change in the demanded AUs is not accompanied by any extension of the service time, then the average service intensity $\mu_{i,k}$ in each area, pre- and postthreshold, is identical. Therefore

$$\begin{aligned} \forall_{0 \leq k \leq q_i} \lambda_{i,k} &= \lambda_i, \\ \mu_{i,k} &= \mu_i. \end{aligned} \tag{4}$$

Following (4), the intensity of traffic of class i in the prethreshold area and in each postthreshold area is identical:

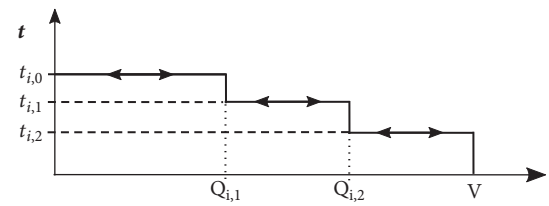


Figure 2: MTS for a single class and for $q_i = 2$.

$$\forall_{0 \leq k \leq q_i} A_{i,Er,k} = \frac{\lambda_i}{\mu_i} = A_{i,Er}. \tag{5}$$

The symbol Er in the lower index indicates that traffic under consideration is Erlang traffic.

The operation of MTS with the example of one call class for two thresholds is shown in Figure 2. In the prethreshold area ($0 \leq n \leq Q_{i,1}$), the CAC function allocates to each call (of class i) $t_{i,0}$ AUs. An increase in the load of the system, i.e., the crossing of the threshold $Q_{i,1}$ results in a transition to the first postthreshold area ($Q_{i,1} < n \leq Q_{i,2}$), in which the number of allocated AUs for each call of class i is $t_{i,1}$. A transition to the second postthreshold area ($Q_{i,2} < n \leq V$) is followed by a decrease in the number of allocated AUs to the value $t_{i,2}$. The occupancy distribution in MTS can be determined by the following recurrence formula:

$$n\left[P_n\right]_V = \sum_{i=1}^{m} \sum_{k=0}^{q_i} A_{i,Er} t_{i,k} \delta_{i,k}\left(n - t_{i,k}\right)\left[P_{n-t_{i,k}}\right]_V, \tag{6}$$

where $[P_n]_V$ is occupancy probability n AUs in FAR with the capacity V AUs and $\delta_{i,k}(n)$ is conditional probability of transition for a stream of class i that can be determined in the following way:

$$\delta_{i,k}(n) = \begin{cases} 1 & \text{for } Q_{i,k} < n \leq Q_{i,k+1}, \\ 0 & \text{for the remaining } n. \end{cases} \tag{7}$$

The assumption in Formula (7) is that $Q_{i,0} = 0$. Notice that the conditional transitional probability $\delta_{i,k}(n)$ is equal to unity only in such a load area in which the number of allocated AUs is equal to $t_{i,k}$.

The blocking probability in MTS can be determined on the basis of Formula (8):

$$E_i = \sum_{k=0}^{q_i} E_{i,k}, \tag{8}$$

where $E_{i,k}$ is the blocking probability of class i in the postthreshold area k:

$$E_{i,k} = \begin{cases} 0 & \text{for } \begin{cases} V - t_{i,k} \geq Q_{i,k+1}, \\ V - t_{i,k} > Q_{i,k}, \end{cases} \\ \displaystyle\sum_{n=V-t_{i,k}+1}^{Q_{i,k+1}} \left[P_n\right]_V & \text{for } \begin{cases} V - t_{i,k} < Q_{i,k+1}, \\ V - t_{i,k} > Q_{i,k}, \end{cases} \\ \displaystyle\sum_{n=Q_k+1}^{Q_{i,k+1}} \left[P_n\right]_V & \text{for } \begin{cases} V - t_{i,k} < Q_{i,k+1}, \\ V - t_{i,k} \leq Q_{i,k}. \end{cases} \end{cases} \tag{9}$$

On the basis of (9) it is possible to verify that the blocking probability in a given postthreshold area k depends on the position of a given reference state $V-t_{i,k}$ in relation to the pair of thresholds $Q_{i,k}$ and $Q_{i,k+1}$ that limit the threshold area k. The assumption in our further considerations is that blocking of the system can occur only in the oldest post-threshold area q_i. This means that all reference states $V - t_{i,k}$, for $k \neq q_i$, are located above the threshold $Q_{i,k+1}$, whereas the threshold Q_{i,q_i} is selected in such a way that the conditions $Q_{i,q} < V-t_{i,q_i} < V$ are fulfilled. In this particular case the blocking probability in the threshold system is defined by the following formula:

$$E_i = E_{i,q_i} = \sum_{n=V-t_{i,q_i}+1}^{V} [P_n]_V . \tag{10}$$

It should be stressed that such a choice as to the thresholds, where blocking state is possible only in the oldest postthreshold area, is in compliance with the very idea of the introduction of threshold mechanism; i.e., to avoid the blocking phenomenon, the system decreases bitrates of new calls after successive thresholds are crossed and it is only in the oldest postthreshold area (where the lowest bitrate is allocated to new calls) that a certain number of calls are blocked. Therefore, a selection of thresholds that allows blocking in different postthreshold areas to occur is not a workable choice from the engineering point of view.

If a given traffic class i is a class of streaming traffic that does not undergo threshold compression, then, in line with the adopted notation, one threshold is introduced to this traffic with the value equal to the capacity of FAR, i.e., $Q_{i,1} = V$. This means that the system in all the areas of possible occupancies can be treated as a prethreshold area in which the number of allocated AUs is always constant and is $t_{i,0}$. Hence, streaming traffic of class i can be described by the following conditions:

$$\delta_{i,k}(n) = \begin{cases} 1 & \text{for } k = 0 \text{ and } 0 \le n \le V, \\ 0 & \text{for } k = 1, \end{cases} \tag{11}$$

$$t_{i,k} = \begin{cases} t_{i,0} & \text{for } k = 0, \\ 0 & \text{for } k = 1. \end{cases} \tag{12}$$

If all traffic classes are of streaming nature, then (6) comes in its essence to the well-known recurrence [28, 29]:

$$n[P_n]_V = \sum_{i=1}^{m} A_{i,Er} t_{i,0} \left[P_{n-t_{i,0}}\right]_V . \tag{13}$$

2.3. Model of Resources with Threshold Compression for Adaptive Erlang, Engset, and Pascal Traffic. Consider a FAR in which the capacity is V AUs and to which m traffic classes from the set M are offered. The assumption is that m_{Er} classes that belong to the set M_{Er} are of Erlang type, m_{En} classes that belong to the set M_{En} are Engset traffic, and m_{Pa} classes that belong to the set M_{Pa} are Pascal traffic, while $m_{Er} + m_{En} + m_{Pa} = m$ and $M_{Er} \cup M_{En} \cup M_{Pa} = M$. Engset and Pascal traffic are state-dependent and depend on the state of the system, in particular on the number of Engset and Pascal calls that are already being serviced [31]:

$$A_{j,En}(n) = \alpha_{j,En}\left[S_{j,En} - y_{j,En}(n)\right], \quad \text{where } j \in M_{En}, \tag{14}$$

$$A_{l,Pa}(n) = \alpha_{l,Pa}\left[S_{l,Pa} + y_{l,Pa}(n)\right], \quad \text{where } l \in M_{Pa}, \tag{15}$$

where

(i) $A_{j,En}(n)$ is average intensity of Engset traffic of class j in FAR that is in the occupancy state n AUs,

(ii) $A_{l,Pa}(n)$ is average intensity of Pascal traffic of class l in FAR that is in the occupancy state n AUs,

(iii) $\alpha_{j,En}$ is average intensity of Engset traffic of class j generated by one free source:

$$\alpha_{j,En} = \frac{\gamma_{j,n}}{\mu_{j,n}}, \tag{16}$$

(iv) $\gamma_{j,n}$ is average intensity of Engset calls of class j, generated by one free source,

(v) $\mu_{j,n}$ is average service intensity for Engset calls of class j,

(vi) $\alpha_{l,Pa}$ is average intensity of Pascal traffic of class l generated by one free source:

$$\alpha_{l,Pa} = \frac{\gamma_{l,Pa}}{\mu_{l,Pa}}, \tag{17}$$

(vii) $\gamma_{l,Pa}$ is average intensity of Pascal calls of class l generated by one free source,

(viii) $\mu_{l,Pa}$ is average service intensity for Pascal calls of class l,

(ix) $S_{j,En}$ is the number of Engset traffic sources of class j,

(x) $S_{l,Pa}$ is the number of Pascal traffic sources of class l.

The parameters $y_{j,En}(n)$ and $y_{l,Pa}(n)$ are the average values of the number of serviced Engset calls of class j and Pascal calls of class l in FAR that are in the occupancy state n AUs. The method for a determination of these parameters will be given further on in the section. Let us assume now that each class of Engset and Pascal traffic will be assigned its own set of thresholds and that traffic is of adaptive nature. The occupancy distribution in MTS to which a mixture of Erlang, Engset, and Pascal traffic is offered can be determined on the basis of the following recurrence formula [51]:

$$
\begin{aligned}
n\left[P_n\right]_V = & \sum_{i\in M_{Er}}\sum_{k=0}^{q_j} A_{i,Er}t_{i,k}\delta_{i,k}\left(n-t_{i,k}\right)\left[P_{n-t_{i,k}}\right]_V \\
& + \sum_{j\in M_{En}}\sum_{k=0}^{q_j}\alpha_{j,En}\left[S_{j,En}-y_{j,En,k}\left(n-t_{j,k}\right)\right] \\
& \cdot t_{j,k}\delta_{j,k}\left(n-t_{j,k}\right)\left[P_{n-t_{j,k}}\right]_V + \sum_{l\in M_{Pa}}\sum_{k=0}^{q_l}\alpha_{l,Pa} \\
& \cdot \left[S_{l,Pa}+y_{l,Pa,k}\left(n-t_{l,k}\right)\right]t_{l,k}\delta_{l,k}\left(n-t_{l,k}\right) \\
& \cdot \left[P_{n-t_{l,k}}\right]_V,
\end{aligned}
\tag{18}
$$

where the parameters $y_{j,En,k}$ and $y_{l,Pa,k}(n)$ are the average numbers of serviced Engset calls of class j and Pascal calls of class l in FAR that are in the occupancy state n AUs, whereas n belongs to the prethreshold area ($k = 0$) or postthreshold area ($0 < k \le q_j, 0 < k \le q_l$). These parameters can be determined on the basis of the following formula:

$$
\begin{aligned}
& y_{j,En,k}(n) \\
& = \frac{\alpha_{j,En}\left[S_{j,En}-y_{j,En,k}\left(n-t_{j,k}\right)\right]\left[P_{n-t_{j,k}}\right]_V}{\left[P_n\right]_V},
\end{aligned}
\tag{19}
$$

$$
y_{l,Pa,k}(n) = \frac{\alpha_{l,Pa}\left[S_{l,Pa}+y_{l,Pa,k}\left(n-t_{l,k}\right)\right]\left[P_{n-t_{l,k}}\right]_V}{\left[P_n\right]_V}.
\tag{20}
$$

Formulas (19) and (20) define, respectively, the parameters $y_{j,En,k}(n)$ and $y_{l,Pa,k}(n)$ for the following ranges of the variable n: $Q_{j,k}+t_{j,k} < n \le Q_{j,k+1}$; $Q_{l,k}+t_{l,k} < n \le Q_{l,k+1}$ [51].

Assuming, exactly as in the case of MTS with Erlang traffic, that blocking of the system can occur only in the oldest postthreshold area, the blocking probability for calls of individual types can be expressed by Formula (10) that in the adopted notation for particular traffic types can be written as follows:

$$
E_{i,Er} = \sum_{n=V-t_{i,q_i}+1}^{V}\left[P_n\right]_V \quad \text{for } i \in M_{Er},
\tag{21}
$$

$$
E_{j,En} = \sum_{n=V-t_{j,q_j}+1}^{V}\left[P_n\right]_V \quad \text{for } j \in M_{En},
\tag{22}
$$

$$
E_{l,Pa} = \sum_{n=V-t_{l,q_l}+1}^{V}\left[P_n\right]_V \quad \text{for } l \in M_{Pa}.
\tag{23}
$$

In the occupancy distribution FAR with adaptive traffic executed in the threshold manner (18), there are the parameters $y_{j,En,k}(n)$ and $y_{l,Pa,k}(n)$ that can, in turn, be determined on the basis of the occupancy distribution (Formulas (19) and (20)). Therefore, to determine the distribution (18) it is necessary to construct an iterative algorithm in which in each iteration step the approximate occupancy distribution can be determined on the basis of the values of the average number of serviced calls of Engset and Pascal classes, determined in the preceding iteration step. A general method for a construction of these algorithms is proposed in [31, 34].

3. Modelling of Primary Resources with Adaptive Traffic and Threshold Compression

Figure 3 shows a general diagram of a multiservice traffic overflow system. The system of primary resources (PRs) is composed of r FARs, from which each can be considered as MTS with Erlang, Engset, and Pascal adaptive traffic.

Each PR s ($0 < s \le r$) has the capacity $V^{(s)}$ expressed in AUs. The primary resources s are offered $m^{(s)}$ traffic classes from the set $M^{(s)}$, where $m_{Er}^{(s)}$ classes that belong to the set $M_{Er}^{(s)}$ are Erlang, $m_{Er}^{(s)}$ classes that belong to the set $M_{En}^{(s)}$ are Engset, and $m_{Pa}^{(s)}$ classes that belong to the set $M_{Pa}^{(s)}$ are Pascal traffic sources, while $m_{Er}^{(s)} + m_{En}^{(s)} + m_{Pa}^{(s)} = m^{(s)}$ and $M_{Er}^{(s)} \cup M_{En}^{(s)} \cup M_{Pa}^{(s)} = M^{(s)}$. In Figure 3, the following notation for the respective traffic intensities is adopted:

(i) $A_{c,X}^{(s)}$ is average intensity of traffic of class c ($c \in M^{(s)}$) of type X ($X = Er \mid En \mid Pa$) offered to PR no. s,

(ii) $R_{c,X}^{(s)}$ is average intensity of traffic of class c ($c \in M^{(s)}$) that overflows from PR no. s, with the assumption that traffic of class c offered to primary resources s is of type X ($X = Er \mid En \mid Pa$).

In each PR s, for each call class c, an individual set $q_c^{(s)}$ of thresholds $\{Q_{c,1}^{(s)}, Q_{c,1}^{(s)}, \ldots, Q_{c,1}^{(s)}\}$ and an individual set $q_c^{(s)} + 1$ of demands allocated in appropriate load areas $\{t_{c,0}^{(s)}, t_{c,1}^{(s)}, \ldots, t_{c,q_c}^{(s)},\}$ are defined. With these assumptions and on the basis of the model presented in Section 2.3, Formulas (18)–(23) can be applied to determine blocking probabilities for individual call classes in each primary resource s. The blocking probability for calls of class c ($c \in M^{(s)}$) of type X ($X = Er \mid En \mid Pa$) can be written on the basis of (21)–(23) as follows:

$$
E_{c,X}^{(s)} = \sum_{n=V^{(s)}-t_{c,q_c}^{(s)}+1}^{V^{(s)}}\left[P_n\right]_{V^{(s)}} \quad \text{for } c \in M_X^{(s)}.
\tag{24}
$$

3.1. Decomposition of Primary Resources. In multiservice models that are based on a description of overflow traffic with two moments, the average value and variance, each PR s undergoes decomposition into $m^{(s)}$ fictitious primary resources (FPR s_c) [24], each with the capacity $v_{c,X}^{(s)}$. Then, each FPR s_c is replaced by equivalent fictitious primary resources (EFPR s_c). Each EFPR s_c can be characterized by equivalent capacity $v_{c,X}^{*(s)}$ and offered, equivalent, Erlang traffic with the intensity $A_{c,X}^{*(s)}$, while these parameters are selected in such a way that traffic that overflows from EFPR s_c has exactly the same parameters (average value and variance) as traffic that overflows from FPR s_c.

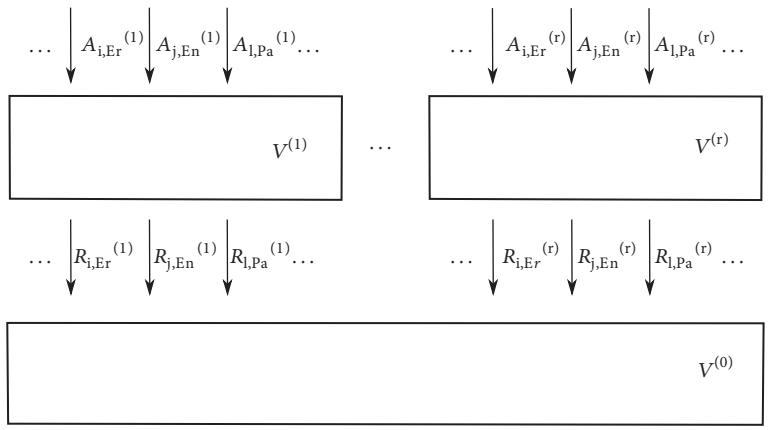


Figure 3: Multiservice overflow traffic system.

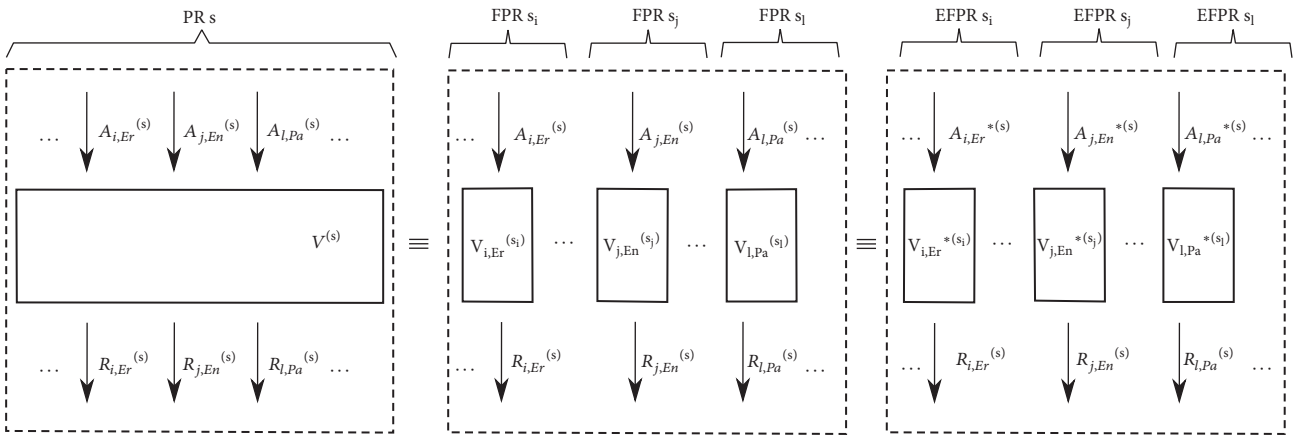


Figure 4: Decomposition of PR s.

3.2. Determination of the Parameters of FPR s_c. The need for decomposition results from the fact that PR s carries multiservice traffic, and in consequence a direct determination of variance for traffic of individual classes is not possible. Each FPR s_c services exclusively calls of just one class c, which, further on in the article, will allow Riordan formulas [3] to be applied to determine variance of traffic of class c that overflows from FPR s_c. The following assumptions are made to determine the capacity of FPRs:

(1) Blocking probability $E_{c,X}^{(s_c)}$ of calls of class c type X in FPR s with the capacity $v_{c,X}^{(s)}$ is exactly the same as the blocking probability $E_{c,X}^{(s)}$ of calls of this class in PR s with the capacity $V^{(s)}$:

$$E_{c,X}^{(s_c)} = E_{c,X}^{(s)} \quad \text{for } c \in M_X^{(s)}. \tag{25}$$

(2) Traffic in FPR s_c does not undergo the threshold compression mechanism.

Figure 4 shows the way PR s is decomposed into FPR s_c and EFPR s_c.

The adopted assumptions define the method for a decomposition of the resources. As it is mentioned earlier, on the basis of the model presented in Section 2.3 (Formulas (18)–(23), it is possible to determine blocking probabilities for individual call classes in each of the PRs s. It results from the adoption of Assumption (1) (Formula (25)) that the obtained probabilities are exactly the same as the probabilities in the corresponding FPR. Assumption (2) indicates that there is a possibility to determine the capacity $v_{c,X}^{(s)}$ of decomposed resources on the basis of single-service models FAR for Erlang, Engset, and Pascal traffic, in which it is possible to determine the blocking probability as the occupancy probability of all AUs:

$$E_{c,X}^{(s_c)} = E_{c,X}^{(s)} = \left[P_{v_{c,X}^{(s)}}\right]_{v_{c,X}^{(s)}} \quad \text{for } c \in M_X^{(s)}. \tag{26}$$

Thus, in the case of Erlang traffic, the capacity of the fictitious primary resources can be determined on the basis of Erlang B Formula:

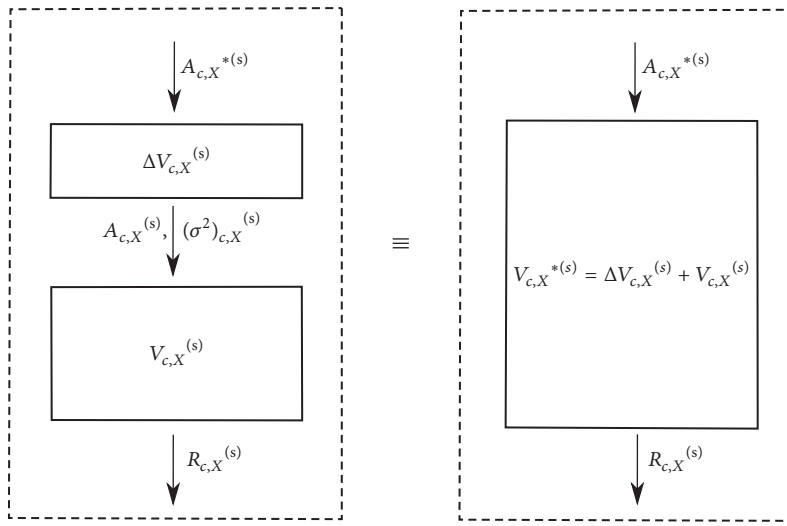


Figure 5: Diagram of the replacement of FPR s_c by EFPR s_c.

$$E_{i,Er}^{(s_i)} = \left[P_{v_{i,Er}^{(s)}}\right]_{v_{i,Er}^{(s)}} = E_{v_{i,Er}^{(s)}}\left(A_{i,Er}^{(s)}\right) = \frac{\left(A_{i,Er}^{(s)}\right)^{v_{i,Er}^{(s)}} / \left(v_{i,Er}^{(s)}\right)!}{\sum_{n=0}^{v_{i,Er}^{(s)}} \left(\left(A_{i,Er}^{(s)}\right)^n / n!\right)}. \tag{27}$$

In Formula (27), the parameters $E_{i,Er}^{(s_i)}$ and $A_{i,Er}^{(s)}$ are known; hence, on their basis, it is possible to determine the parameter $v_{i,Er}^{(s)}$. In a similar way, the capacities of fictitious resources to which single-service Engset and Pascal traffic is offered can be determined on the basis of appropriate formulas that determine the blocking probability in the Engset and Pascal model:

$$E_{j,En}^{(s_j)} = \left[P_{v_{j,En}^{(s)}}\right]_{v_{j,En}^{(s)}} = \frac{\binom{S_{j,En}^{(s)}}{v_{j,En}^{(s)}} \left(\alpha_{j,En}^{(s)}\right)^{v_{j,En}^{(s)}}}{\sum_{n=0}^{v_{j,En}^{(s)}} \binom{S_{j,En}^{(s)}}{n} \left(\alpha_{j,En}^{(s)}\right)^{v_{j,En}^{(s)}}}, \tag{28}$$

$$E_{l,Pa}^{(s_l)} = \left[P_{v_{l,Pa}^{(s)}}\right]_{v_{l,Pa}^{(s)}} = \frac{\binom{-S_{l,Pa}^{(s)}}{v_{l,Pa}^{(s)}} \left(-\alpha_{l,Pa}^{(s)}\right)^{v_{l,Pa}^{(s)}}}{\sum_{n=0}^{v_{l,Pa}^{(s)}} \binom{-S_{l,Pa}^{(s)}}{n} \left(-\alpha_{l,Pa}^{(s)}\right)^{v_{l,Pa}^{(s)}}}. \tag{29}$$

Note that Formulas (27)–(29) apply to single-service systems in which admission of a new call means the occupation of one AU.

3.3. Determination of the Parameters of EFPR s_c. The next element in the decomposition of PR s is the replacement of FPR s_c by EFPR s_c. The need for such a replacement is related to a possibility to determine, further on in this article, the variance of traffic that overflows from EFPR s_c with the help of Riordan formulas [3]. Since these formulas make it possible to determine variance exclusively in the case where offered traffic is Erlang traffic, then a replacement of FPR s_c by EFPR s_c demands a replacement of Engset and Pascal traffic by equivalent Erlang traffic. Let us denote the variance of traffic of class c type X by $(\sigma_\Delta^2)_{c,X}^{(s)}$. Equivalent Erlang traffic of $A_{j,En}^{*(s)}$ and $A_{l,Pa}^{*(s)}$ is such traffic offered to certain fictitious, additional resources with the capacities $\Delta v_{j,En}^{(s)}$ and $\Delta v_{l,Pa}^{(s)}$ that traffic that overflows from these resources is equal, with respect to the average value and variance, to Engset $(A_{j,En}^{(s)}, (\sigma_\Delta^2)_{j,En}^{(s)})$ and Pascal traffic $(A_{l,Pa}^{(s)}, (\sigma_\Delta^2)_{l,Pa}^{(s)})$, respectively. Note that, in the notation of equivalent traffic, e.g., in $A_{j,En}^{*(s)}$, the symbol En that indicates the primary nature of this traffic has been retained in the lower index. The "asterisk" introduced to the upper index means that this traffic is already equivalent Erlang traffic. In the case where primary traffic is Erlang traffic, we have $A_{i,Er}^{*(s)} = A_{i,Er}^{(s)}$, $(\sigma_\Delta^2)_{i,Er}^{(s)} = A_{i,Er}^{(s)}$, and $\Delta v_{i,Er}^{(s)} = 0$.

The EFPR parameters s_c can be determined on the basis of the equivalent random theory method (ERT) [2, 3] that, for Engset traffic, is described in [8]. The method will be generalized to include the particular case of Pascal traffic. The diagram of traffic replacement from Engset and Pascal traffic to equivalent Erlang traffic is presented in Figure 5. Average values and variances for Engset and Pascal traffic can be determined on the basis of the following dependencies:

(i) Erlang traffic:

$$A_{i,Er}^{(s)}, \left(\sigma_\Delta^2\right)_{i,\mathrm{Er}}^{(s)} = A_{i,Er}^{(s)}, \tag{30}$$

(ii) Engset traffic:

$$A_{j,En}^{(s)} = S_{j,En}^{(s)} \frac{\alpha_{j,En}^{(s)}}{1 + \alpha_{j,En}^{(s)}}, \qquad \left(\sigma_\Delta^2\right)_{j,En}^{(s)} = S_{j,En}^{(s)} \frac{\alpha_{j,En}^{(s)}}{\left(1 + \alpha_{j,En}^{(s)}\right)^2}, \tag{31}$$

(iii) Pascal traffic:

$$A_{l,Pa}^{(s)} = S_{l,Pa}^{(s)} \frac{\alpha_{l,Pa}^{(s)}}{1-\alpha_{l,Pa}^{(s)}},$$
$$\left(\sigma_{\Delta}^{2}\right)_{l,Pa}^{(s)} = S_{l,Pa}^{(s)} \frac{\alpha_{l,Pa}^{(s)}}{\left(1-\alpha_{l,Pa}^{(s)}\right)^{2}}. \quad (32)$$

According to the ERT method, the average value R and variance σ^2 of traffic that overflows from PR with the capacity V to which Erlang traffic A is offered can be determined on the basis of Riordan formulas [3]:

$$R = AE_V(A), \quad (33)$$
$$\sigma^2 = R\left(\frac{A}{V+1-A+R}+1-R\right). \quad (34)$$

In the considered traffic replacement diagram (Figure 5), the resources with the capacity $V = \Delta v_{c,X}^{(s)}$ is offered equivalent Erlang traffic with the intensity $A = A_{c,X}^{*(s)}$. Traffic that overflows from these resources has the average value $R = A_{c,X}^{(s)}$ and variance $(\sigma_{\Delta}^2)_{c,X}^{(s)}$. Therefore, in the notation adopted in the article, Formulas (33) and (34) can be written in the following way:

$$A_{c,X}^{(s)} = A_{c,X}^{*(s)} E_{\Delta v_{c,X}^{(s)}}\left(A_{c,X}^{*(s)}\right), \quad (35)$$

$$\left(\sigma_{\Delta}^{2}\right)_{c,X}^{(s)} = A_{c,X}^{(s)}\left(\frac{A_{c,X}^{*(s)}}{\Delta v_{c,X}^{(s)}+1-A_{c,X}^{*(s)}+A_{c,X}^{(s)}}+1-A_{c,X}^{(s)}\right). \quad (36)$$

Formulas (35) and (36) allow the pair of parameters $(A_{c,X}^{*(s)}, \Delta v_{c,X}^{(s)})$ to be determined on the basis of known values of the pair of parameters $(A_{c,X}^{(s)}, (\sigma_{\Delta}^2)_{c,X}^{(s)})$ that, depending on considered type of traffic, are described by Formulas (30)–(32). At this particular point it is worthwhile to note the fact that the diagram of equivalent replacement of FPR s_c by EFPR s_c for Engset traffic offers a solution exclusively for negative values of the capacity $\Delta v_{c,X}^{(s)}$ in Formulas (35) and (36).

Having determined the value of the parameters $(A_{c,X}^{*(s)}, \Delta v_{c,X}^{(s)})$, it is possible to evaluate the capacity of EFPR s_c, i.e., the parameter $v_{c,X}^{*(s)}$ [3]:

$$v_{c,X}^{*(s)} = v_{c,X}^{(s)} + \Delta v_{c,X}^{(s)}. \quad (37)$$

Note that the pair of parameters $(A_{c,X}^{*(s)}, v_{c,X}^{*(s)})$ for each EFPR s_c determines the Erlang model for FAG [81].

3.4. Determination of Parameters of Traffic That Overflows from EFPR s_c. Traffic of class c that overflows from PR s is equivalent to traffic that overflows from EFPR s_c and will be characterized by two parameters: the average value of traffic intensity $R_{c,X}^{(s)}$ and variance $(\sigma^2)_{c,X}^{(s)}$. Since EFPR s_c are determined by a single-service Erlang model, then the parameters $R_{c,X}^{(s)}$ and $(\sigma^2)_{c,X}^{(s)}$ can be determined on the basis of Riordan formulas (33) and (34) that, in the adopted notation, will be written in the following way:

$$R_{c,X}^{(s)} = A_{c,X}^{*(s)} E_{v_{c,X}^{*(s)}}\left(A_{c,X}^{*(s)}\right), \quad (38)$$

$$\left(\sigma^2\right)_{c,X}^{(s)} = R_{c,X}^{(s)}\left(\frac{A_{c,X}^{*(s)}}{v_{c,X}^{*(s)}+1-A_{c,X}^{*(s)}+R_{c,X}^{(s)}}+1-R_{c,X}^{(s)}\right). \quad (39)$$

The peakedness factor $Z_{c,X}^{(s)}$ for traffic of class c that overflows from EFPR s_c is defined as the ratio of variance to the average value:

$$Z_{c,X}^{(s)} = \frac{\left(\sigma^2\right)_{c,X}^{(s)}}{R_{c,X}^{(s)}}. \quad (40)$$

Further on in the article, traffic that overflows from EFPR s_c will be characterized by the pair of parameters $(R_{c,X}^{(s)}$ i $(\sigma^2)_{c,X}^{(s)})$.

4. Modeling of Secondary Resources with Adaptive Traffic and Threshold Compression

The parameters $R_{c,X}^{(s)}$ and $(\sigma^2)_{c,X}^{(s)}$ of traffic that overflows from individual EFPR s_c are simultaneously the parameters for traffic offered to SR with the capacity $V^{(0)}$ AUs. In SR, traffic can also undergo the threshold compression mechanism. This means that, in exactly the same way as in the case of PR, for each call class c $(0 < c \le m)$ the set $q_c^{(0)}$ of thresholds $\{Q_{c,1}, Q_{c,2}, \ldots, Q_{c,q_c}\}$ is introduced individually, where the occupancy area that satisfies the condition $0 < n \le Q_{c,1}$ is the prethreshold area, whereas the occupancy area that satisfies the condition $Q_{c,k} \le Q_{c,k+1}$ is called the postthreshold area k, where $1 \le k \le q_c^{(0)}$. In each postthreshold area k, the number of AUs allocated to service calls of class c is $t_{c,k}^{(0)}$, while the value of this parameter in the prethreshold area is $t_{c,0}^{(0)}$.

4.1. Determination of the Occupancy Distribution in SR. To model the system of secondary resources in this article, Hayward approach [10] is used. The approach is based on a division of SR parameters (traffic intensity and capacity) by the peakedness factor of offered traffic. Such an approach is used in [24] to model secondary resources composed of a single FAR and servicing a mixture of different classes of overflow traffic. In this section, Hayward approach is used to describe the characteristics of SR that is composed of a single MTS. As a result, the occupancy distribution and blocking probability in SR can be written, using the adopted notation, as follows:

$$n\left[P_n\right]_{V^{(0)}/Z^{(0)}} = \sum_{s=1}^{r}\sum_{c=1}^{m^{(s)}}\sum_{k=0}^{q_c^{(0)}} \frac{R_{c,X}^{(s)}}{Z_{c,X}^{(s)}} t_{c,k}^{(0)} \delta_{c,k}^{(0)}\left(n-t_{c,k}^{(0)}\right)\left[P_{n-t_{c,k}^{(0)}}\right]_{V^{(0)}/Z^{(0)}}, \quad (41)$$

where $\delta_{c,k}^{(0)}(n)$ is the conditional transition probability for a stream of class c that in the case of the distribution (41) can be determined in the following way:

$$\delta_{c,k}^{(0)}(n) = \begin{cases} 1 & \text{for } \dfrac{Q_{i,k}}{Z^{(0)}} < n \le \dfrac{Q_{i,k+1}}{Z^{(0)}}, \\ 0 & \text{for the remaining } n. \end{cases} \tag{42}$$

In Formula (41), the peakedness factors $Z_{c,X}^{(s)}$ of traffic that overflows from PRs are defined by Formulas (38)–(40). The parameter $Z^{(0)}$ is the so-called aggregate peakedness factor. The introduction of this approach results from a necessity, in compliance with Hayward's concept, to normalize the total capacity of the system of secondary resources $V^{(0)}/Z^{(0)}$ to which an appropriate mixture of traffic, which overflows from each of the r of primary resources, is offered. The aggregate peakedness factor can be determined approximately on the basis of the weighted mean of the peakedness factor of an individual traffic classes offered to SR [24]:

$$Z^{(0)} = \frac{\sum_{s=1}^{r} \sum_{c=1}^{m^{(s)}} \left(\sigma^2\right)_{c,X}^{(s)}}{\sum_{s=1}^{r} \sum_{c=1}^{m^{(s)}} R_{c,X}^{(s)}}. \tag{43}$$

After a determination of the occupancy distribution in SR, the blocking probability for each class of calls in secondary resources can be determined:

$$\left(E_{c,X}^{(s)}\right)^{(0)} = \sum_{n=V^{(0)}/Z^{(0)}-t_{c,q_c}^{(0)}+1}^{V^{(0)}} [P_n]_{V^{(0)}/Z^{(0)}}. \tag{44}$$

4.2. Method of Modelling SR. The models presented in Sections 3 and 4 allow a method for a calculation of the blocking probability and other characteristics in a multiservice overflow system with adaptive traffic in primary and secondary resources to be defined. The method can be presented in the following steps:

(1) Determination of occupancy distributions $[P_n]_{V^{(s)}}$ (Formula (18)) and the blocking probability $E_{c,X}^{(s)}$ (Formula (24)) for traffic streams of all classes in each PR s, $1 \le s \le r$.

(2) Determination of the capacity $v_{c,X}^{(s)}$ for each FPR s_c (Formulas (27)–(29)).

(3) Determination of the variance $(\sigma_\Delta^2)_{c,X}^{(s)}$ of each Erlang, Engset, and Pascal traffic offered to PR s, $1 \le s \le r$ (Formulas (30)–(32)).

(4) Determination of the parameters of each of EFPR s_c, i.e., the equivalent intensity of Erlang traffic $A_{c,X}^{*(s)}$ and the equivalent capacity $v_{c,X}^{*(s)}$ (Formulas (35)–(37)).

(5) Determination of the parameters of overflow traffic: mean value of traffic intensity $R_{c,X}^{(s)}$ and variance $(\sigma^2)_{c,X}^{(s)}$ (Formulas (38)–(39)) for traffic of all classes that overflows from the system of primary resources.

(6) Determination of the occupancy distribution $[P_n]_{V^{(0)}/Z^{(0)}}$ (Formula (41)) and the aggregate peakedness factor $Z^{(0)}$ in SR (Formula (43)).

(7) Determination of the blocking probability for traffic streams of all classes offered to SR (Formula (44)).

4.3. Comment. It is possible to determine on the basis of the blocking probability other important QoS characteristics, such as the call loss probability. The loss probability in an overflow system for Erlang traffic is equal to the blocking probability. It has been proved in traffic theory, e.g., [13], that in a single-service Engset (Pascal) model the call loss probability is equal to the blocking probability in a system with identical capacity in which the number of traffic sources has been decreased (increased) by one source. This approach can be then used to approximately determine the call loss probability for Engset and Pascal traffic classes in multiservice systems. Therefore, the loss probability for Engset (Pascal) traffic classes can be determined on the basis of the blocking probability, assuming that the number of traffic sources of each Engset (Pascal) traffic class has been decreased (increased) by one source.

5. Numerical Examples

The proposed model of multiservice overflow system with adaptive traffic is an approximate one. In order to determine its accuracy, the results of analytical modelling have been compared with the simulation data obtained for the system described in Table 1.

The system under consideration is characterised in Table 1 by specifying the number of primary resources and their capacity, the capacity of the secondary resources, and the number of AUs required by particular traffic classes in prethreshold and postthreshold areas. It was assumed that each of primary resource was offered traffic of various classess in the following proportions: $A_{1,0,s}t_{1,0,s} : A_{2,0,s}t_{2,0,s} : \cdots = 1 : 1 : \ldots$. We can notice that the PR no. 1 was offered three traffic classes (flows, services), with the demands equal to 1, 6, and 12 AUs, respectively. In the case of the PR no. 2, only two traffic classes were offered, demanding 3 and 11 AUs. The PR no. 3 and PR no. 4 were dedicated for servicing only a single traffic class. Such distribution of demands for particular resources can be treated as an example of allocating of separated resources (virtual resources, slices) to particular demands. The flows that cannot be serviced by dedicated primary resources are offered to the common secondary resources. In both types of resources the flows undergo threshold compression, according to the parameters specified in Table 1. The volume of resources admitted for flows in particular load areas were different in primary and secondary resources.

The results of blocking probabilities for all traffic classes in both primary and secondary resources are presented in Figures 6–9 and 10, respectively. In the case of the secondary resources, the results of blocking probability for traffic classes nos. 3, 5, and 6 were the same since all these classes were admitted the same amount of AUs in the last threshold

Table 1: Description of the system with threshold compression in primary and secondary resources. The secondary resources are denoted by $s = 0$.

s	$V^{(s)}$	k	c	$Q_{c,k}^{(s)}$	$t_{c,k}^{(s)}$	typ	$S_{c,k}^{(s)}$
		0	1	0	1	Er	–
		0	2	0	6	En	10
		0	3	0	12	Pa	10
		1	2	36	3	En	10
		1	3	32	8	Pa	10
1	48	2	3	42	6	Pa	10
		0	4	0	3	Er	–
		0	5	0	11	En	12
		1	4	30	2	Er	–
2	50	1	5	24	6	En	12
		0	6	0	15	En	20
		1	6	30	12	En	20
3	100	2	6	50	10	En	20
		0	7	0	17	Pa	10
4	120	1	7	50	12	Pa	10
		0	1	0	1	Er	–
		0	2	0	6	En	10
		0	3	0	12	Pa	10
		0	4	0	3	Er	–
		0	5	0	11	En	12
		0	6	0	15	En	20
		0	7	0	17	Pa	10
		1	2	20	3	En	10
		1	3	20	10	Pa	10
		1	4	30	2	Er	–
		1	5	30	5	En	12
		1	6	20	10	En	20
		1	7	20	12	Pa	10
		2	3	30	5	Pa	10
		2	6	30	5	En	20
0	100	2	7	30	10	Pa	10

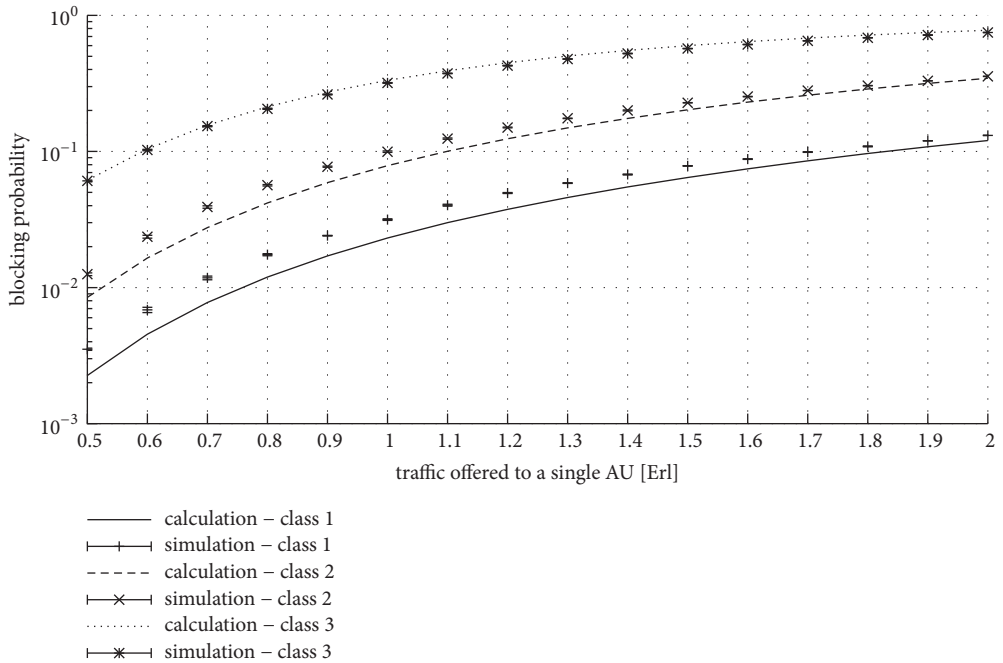


Figure 6: Blocking probability in primary resources (slice) no. 1 for traffic classes nos. 1, 2, and 3 from Table 1. Traffic offered per single AU calculated for initial value of demanded AUs (before compression).

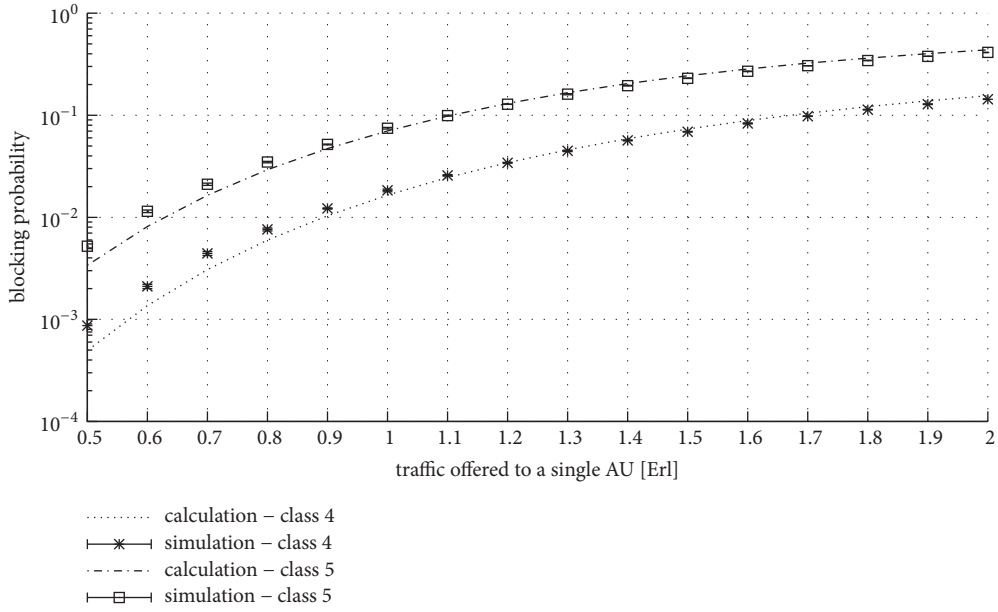


Figure 7: Blocking probability in primary resources (slice) no. 2 for traffic classes nos. 4 and 5 from Table 1. Traffic offered per single AU calculated for initial value of demanded AUs (before compression).

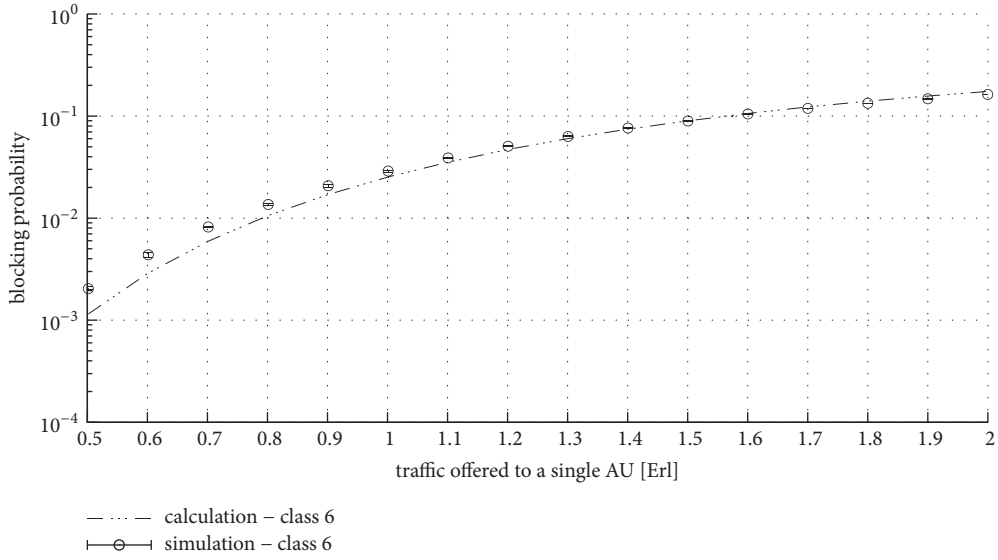


Figure 8: Blocking probability in primary resources (slice) no. 3 for traffic class no. 6 from Table 1. Traffic offered per single AU calculated for initial value of demanded AUs (before compression).

(in the area of the highest load). In order to increase the readability of the Figure 10, the results for class 4 were omitted.

The results presented in Figures 6–10 indicate good accuracy of the proposed method, for both primary and secondary resources. In the proposed method, the errors within the area of lower loads mainly result from the fact that in Formula (41) the coefficient $V^{(0)}/Z^{(0)}$ can take on noninteger values. In such a case, linear interpolation between the values $\lfloor V^{(0)}/Z^{(0)} \rfloor$ and $\lceil V^{(0)}/Z^{(0)} \rceil$ is applied in the model and this approximation is largely responsible for the decrease in accuracy of the proposed method in the area of low losses. The problem of noninteger values of the coefficients $V^{(0)}/Z^{(0)}$ in generalized Hayward's formula is discussed in [62].

6. Conclusions

This article proposes a model of multiservice traffic overflow system with adaptive traffic that undergoes the threshold compression mechanism in both primary and secondary resources. Both primary and secondary resources can be physical resources as well as virtual resources (e.g., slices).

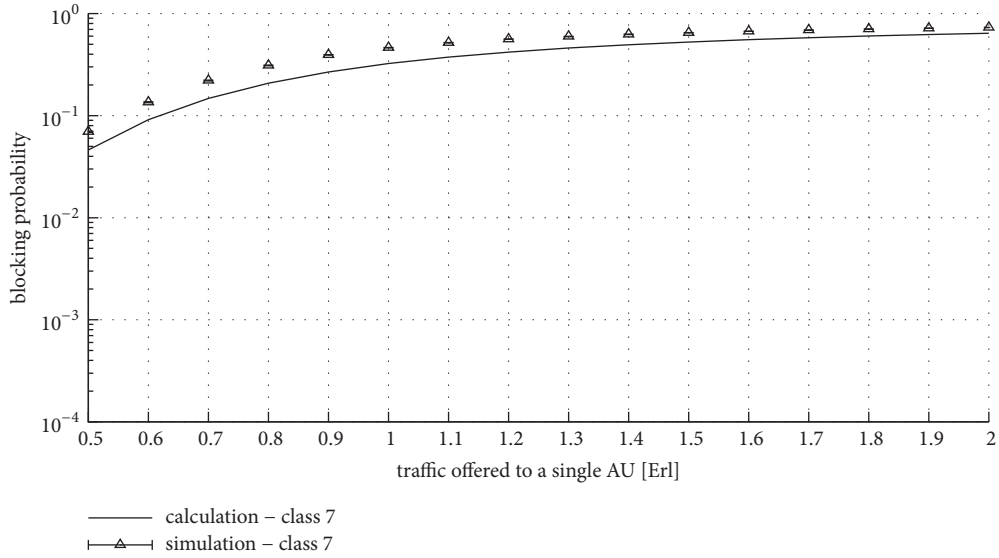


Figure 9: Blocking probability in primary resources (slice) no. 4 for traffic class no. 7 from Table 1. Traffic offered per single AU calculated for initial value of demanded AUs (before compression).

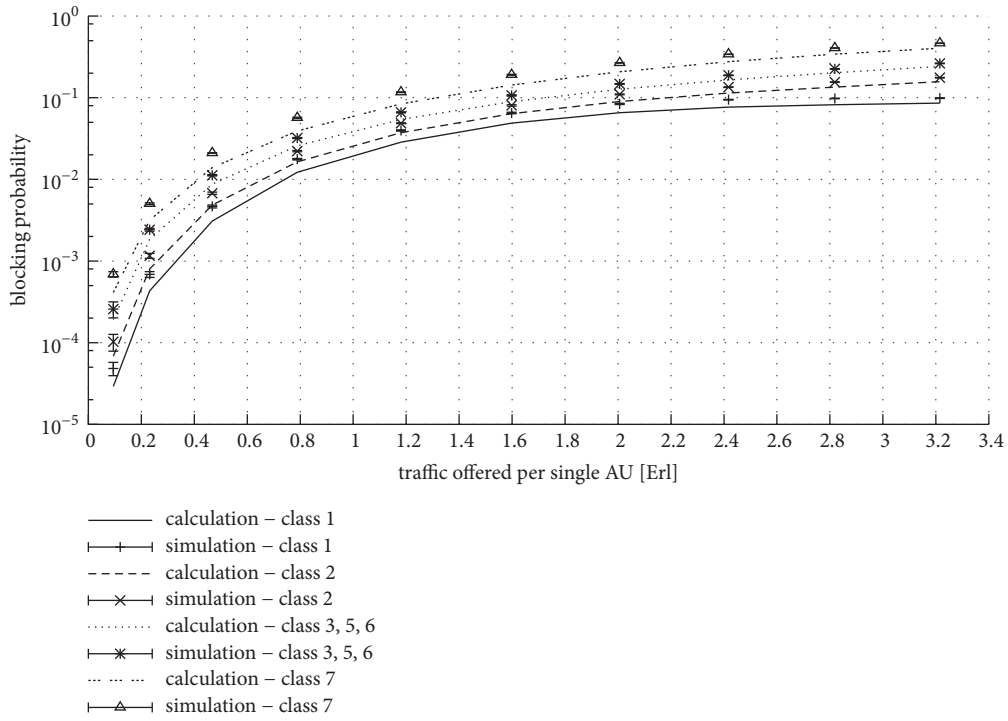


Figure 10: Blocking probability in secondary resources (slice) no. 0 for traffic classes nos. 1, 2, 3, 5, 6, and 7 from Table 1. Traffic offered per single AU calculated for initial value of demanded AUs (before compression).

The model takes into consideration three possible types of traffic: Erlang, Engset, and Pascal traffic. The model is based on a generalization of Hayward's concept and its application to model systems with adaptive traffic with threshold compression. The proposed model is an approximate model and therefore the results of the analytical modeling are compared with the results of the simulation experiments. The results obtained from the comparison prove good accuracy of the proposed model that is independent of both the number and values of introduced thresholds and of the number of classes and the type of offered traffic.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work is supported by the Polish Ministry of Science and Higher Education, Grant 08/82/DSPB/8224.

References

[1] M. Stasiak, M. Głąbowski, A. Wis'niewski, and P. Zwierzykowski, *Modeling and Dimensioning of Mobile Networks*, Wiley, 2011.

[2] G. Bretschneider, "Die berechnung von leitungsgruppen für überfliessenden verkehr in fernsprechwählanlagen," *Nachrichtentechnische Zeitung*, no. 11, pp. 533–540, 1956.

[3] R. I. Wilkinson, "Theories of toll traffic engineering in the USA," *Bell System Technical Journal*, vol. 40, pp. 421–514, 1956.

[4] Y. Rapp, "Planning of junction network in a multi-exchange area," in *Proceedings of the 4th International Teletraffic Congress*, vol. 1964, London, UK.

[5] B. Wallström, "A distribution model for telephone traffic with varying call intensity, including overflow traffic," *Ericsson Technics*, vol. 20, no. 2, pp. 183–202, 1964.

[6] U. Herzog, "Die exakte Berechnung des Streuwertes von überlaufverkehrhinter Koppelanordnungen beliebiger Stufenzahl mit vollkommener bzw. unvollkommener Erreichbarkeit," *International Journal of Electronics And Communications*, vol. 20, no. 3, 1966.

[7] R. Schehrer, "On the exact calculation of overflow systems," in *Proceedings of the in Proceedings of 6th International Teletraffic Congress*, Munich, Germany, 1970, 1970, https://itc-conference.org/_Resources/Persistent/21594392c496c4c0193b4864ec860502ebdec2bc/schehrer70.pdf.

[8] G. Bretschneider, "Extension of the equivalent random method to smooth traffics," in *Proceedings of the in Proceedings of 7th International Teletraffic Congress*, Stockholm, 1973.

[9] R. Schehrer, "On the calculation of overflow systems with a finite number of sources and full availiable groups," *IEEE Transactions on Communications*, vol. 26, no. 1, pp. 75–82, 1978.

[10] A. A. Fredericks, "Congestion in blocking systems—a simple approximation technique," *Bell Labs Technical Journal*, vol. 59, no. 6, pp. 805–828, 1980.

[11] J. F. Shortle, "An equivalent random method with hyper-exponential service," *Performance Evaluation*, vol. 57, no. 3, pp. 409–422, 2004.

[12] E. W. M. Wong, A. Zalesky, Z. Rosberg, and M. Zukerman, "A new method for approximating blocking probability in overflow loss networks," *Computer Networks*, vol. 51, no. 11, pp. 2958–2975, 2007.

[13] V. Iversen, "Teletraffic engineering handbook," Tech. Rep., Technical University of Denmark, 2010.

[14] C. G. Park and D. H. Han, "Comparisons of loss formulas for a circuit group with overflow traffic," *Journal of Applied Mathematics & Informatics*, vol. 30, no. 1-2, pp. 135–145, 2012.

[15] P. Kühn and M. E. Mashaly, "Multi-server, finite capacity queuing system with mutual overflow," in *Proceedings of the 2nd European Teletraffic Seminar*, M. Fiedler, Ed., Karlskrona, Sweden, 2013.

[16] Y. Chan, J. Guo, E. W. Wong, and M. Zukerman, "Surrogate models for performance evaluation of multi-skill multi-layer overflow loss systems," *Performance Evaluation*, vol. 104, pp. 1–22, 2016.

[17] L.-R. Hu and S. Rappaport, "Personal communications systems using multiple hierarchical cellular overlays," in *Proceedings of the 1994 3rd IEEE International Conference on Universal Personal Communications*, pp. 397–401, San Diego, CA, USA.

[18] X. Lagrange and P. Godlewski, "Performance of a hierarchical cellular network with mobility-dependent hand-over strategies," in *Proceedings of the IEEE 46th Vehicular Technology Conference*, vol. 3, pp. 1868–1872, May 1996.

[19] S.-P. Chung and J.-C. Lee, "Performance analysis and overflowed traffic characterization in multiservice hierarchical wireless networks," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 904–918, 2005.

[20] J. S. Kaufman and K. M. Rege, "Blocking in a shared resource environment with batched Poisson arrival processes," *Performance Evaluation*, vol. 24, no. 4, pp. 249–263, 1996.

[21] I. Moscholios, J. Vardakas, M. Logothetis, and A. Boucouvalas, "Congestion probabilities in a batched Poisson multirate loss model supporting elastic and adaptive traffic," *annals of telecommunications - annales des télécommunications*, vol. 68, no. 5-6, pp. 327–344, 2013.

[22] L. E. N. Delbrouck, "On the steady-state distribution in a service facility carrying mixtures of traffic with different peakedness factors and capacity requirements," *IEEE Transactions on Communications*, vol. 31, no. 11, pp. 1209–1211, 1983.

[23] E. A. van Doorn and F. J. M. Panken, "Blocking probabilities in a loss system with arrivals in geometrically distributed batches and heterogenous service requirements," *IEEE/ACM Transactions on Networking*, vol. 1, no. 6, pp. 664–667, 1993.

[24] M. Głąbowski, K. Kubasik, and M. Stasiak, "Modeling of systems with overflow multi-rate traffic," *Telecommunication Systems*, vol. 37, no. 1–3, pp. 85–96, 2008.

[25] M. Głąbowski, K. Kubasik, and M. Stasiak, "Modelling of systems with overflow multi-rate traffic and finite number of traffic sources," in *Proceedings of the 6th International Symposium Communication Systems, Networks and Digital Signal Processing (CSNDSP '08)*, pp. 196–199, Graz, Austria, July 2008.

[26] Q. Huang, K.-T. Ko, and V. B. Iversen, "Approximation of loss calculation for hierarchical networks with multiservice overflows," *IEEE Transactions on Communications*, vol. 56, no. 3, pp. 466–473, 2008.

[27] A. Lotze, "History and development of grading theory," in *Proceedings of the in Proceedings of 5th International Teletraffic Congress*, pp. 148–161, New York, NY, USA, 1967.

[28] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Transactions on Communications*, vol. 29, no. 10, pp. 1474–1481, 1981.

[29] J. Roberts, "A service system with heterogeneous user requirements — application to multi-service telecommunications systems," in *Proceedings of the Performance of Data Communications Systems and their Applications*, G. Pujolle, Ed., pp. 423–431, Amsterdam, Netherlands, 1981.

[30] M. Głąbowski, "Modelling of state-dependent multirate systems carrying BPP traffic," *Annals of Telecommunications-Annales des Télécommunications*, vol. 63, no. 7-8, pp. 393–407, 2008.

[31] M. Głąbowski, M. Stasiak, and J. Weissenberg, "Properties of Recurrent Equations for the Full-Availability Group with BPP Traffic," *Mathematical Problems in Engineering*, vol. 2012, Article ID 547909, 17 pages, 2012.

[32] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Two-dimensional convolution algorithm for modelling multiservice networks with overflow traffic," *Mathematical Problems in Engineering*, vol. 2013, Article ID 852082, 18 pages, 2013.

[33] M. Głąbowski, S. Hanczewski, and M. Stasiak, "Modelling of Cellular Networks with Traffic Overflow," *Mathematical Problems in Engineering*, vol. 2015, Article ID 286490, 15 pages, 2015.

[34] M. Głąbowski, S. Hanczewski, M. Stasiak, and J. Weissenberg, "Modeling Erlang's Ideal Grading with multirate BPP traffic," *Mathematical Problems in Engineering*, vol. 2012, Article ID 456910, 35 pages, 2012.

[35] Q. Huang, Y.-C. Huang, K.-T. Ko, and V. B. Iversen, "Loss performance modeling for hierarchical heterogeneous wireless networks with speed-sensitive call admission control," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 5, pp. 2209–2223, 2011.

[36] B. M. Bakmaz and M. R. Bakmaz, "Solving some overflow traffic models with changed serving intensities," *AEÜ - International Journal of Electronics and Communications*, vol. 66, no. 1, pp. 80–85, 2012.

[37] M. Wang, S. Li, E. W. M. Wong, and M. Zukerman, "Performance analysis of circuit switched multi-service multi-rate networks with alternative routing," *Journal of Lightwave Technology*, vol. 32, no. 2, Article ID 6658839, pp. 179–200, 2014.

[38] J. Postel, "Transmission Control Protocol," Tech. Rep. RFC 793 (INTERNET STANDARD), Internet Engineering Task Force, 1981, 1981, http://www.ietf.org/rfc/rfc793.txt.

[39] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," Tech. Rep. RFC3550 (INTERNET STANDARD), Internet Engineering Task Force, 2003, 2003, http://www.ietf.org/rfc/rfc3550.txt.

[40] G. M. Stamatelos and V. N. Koukoulidis, "Reservation-based bandwidth allocation in a radio ATM network," *IEEE/ACM Transactions on Networking*, vol. 5, no. 3, pp. 420–428, 1997.

[41] T. Bonald and J. Virtamo, "A recursive formula for multirate systems with elastic traffic," *IEEE Communications Letters*, vol. 9, no. 8, pp. 753–755, 2005.

[42] T. Bonald, A. Proutière, and J. Virtamo, "A queueing analysis of max-min fairness, proportional fairness and balanced fairness," *Queueing Systems*, vol. 53, no. 1, pp. 65–84, 2006.

[43] J.-P. Haddad and R. R. Mazumdar, "Congestion in large balanced multirate networks," *Queueing Systems*, vol. 74, no. 2-3, pp. 333–368, 2013.

[44] T. Bonald and A. Proutiere, "Insensitive bandwidth sharing," in *Proceedings of the GLOBECOM 2002 - IEEE Global Communications Conference*, pp. 2659–2663, Taipei, Taiwan.

[45] S. Rácz, B. P. Gerö, and G. Fodor, "Flow level performance analysis of a multi-service system supporting elastic and adaptive services," *Performance Evaluation*, vol. 49, no. 1-4, pp. 451–469, 2002.

[46] J. S. Kaufman, "Blocking with retrials in a completely shared resource environment," *Performance Evaluation*, vol. 15, no. 2, pp. 99–116, 1992.

[47] J. Kaufman, "Blocking in a completely shared resource environment with state dependent resource and residency requirements," in *Proceedings of the IEEE INFOCOM '92: The Conference on Computer Communications*, vol. 3, pp. 2224–2232, IEEE Computer Society Press, Los Alamitos, CA, USA, May 1992.

[48] I. D. Moscholios, M. D. Logothetis, and G. K. Kokkinakis, "Connection-dependent threshold model: A generalization of the Erlang multiple rate loss model," *Performance Evaluation*, vol. 48, no. 1-4, pp. 177–200, 2002.

[49] I. D. Moscholios, M. D. Logothetis, and P. I. Nikolaropoulos, "Engset multi-rate state-dependent loss models," *Performance Evaluation*, vol. 59, no. 2-3, pp. 247–277, 2005.

[50] I. D. Moscholios, M. D. Logothetis, and G. K. Kokkinakis, "Call-burst blocking of ON-OFF traffic sources with retrials under the complete sharing policy," *Performance Evaluation*, vol. 59, no. 4, pp. 279–312, 2005.

[51] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modeling product-form state-dependent systems with BPP traffic," *Performance Evaluation*, vol. 67, no. 3, pp. 174–197, 2010.

[52] M. Sobieraj, M. Stasiak, J. Weissenberg, and P. Zwierzykowski, "Analytical model of the single threshold mechanism with hysteresis for multi-service networks," *IEICE Transactions on Communications*, vol. 95, no. 1, pp. 120–132, 2012.

[53] M. Sobieraj, M. Stasiak, and P. Zwierzykowski, "Model of the Threshold Mechanism with Double Hysteresis for Multi-service Networks," in *Computer Networks*, vol. 291 of *Communications in Computer and Information Science*, pp. 299–313, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[54] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: Enabling enterprises' own software-defined cellular networks," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 146–153, 2016.

[55] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

[56] P. Rost, A. Banchs, I. Berberana et al., "Mobile network architecture evolution toward 5G," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 84–91, 2016.

[57] X. Wu and Z. Ma, "Modeling and Performance Analysis of Cellular and Device-to-Device Heterogeneous Networks," in *Proceedings of the 2017 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, Singapore, Singapore, December 2017.

[58] M. Stasiak, "Queuing systems for the internet," *IEICE Transactions on Communications*, vol. 99, no. 6, pp. 1234–1242, 2016.

[59] M. Głąbowski, S. Hanczewski, and M. Stasiak, "Modelling load balancing mechanisms in self-optimising 4G mobile networks with elastic and adaptive traffic," *IEICE Transactions on Communications*, vol. E99B, no. 8, pp. 1718–1726, 2016.

[60] S. Hanczewski, M. Stasiak, and P. Zwierzykowski, "Modelling of the access part of a multi-service mobile network with service priorities," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, 2015.

[61] M. Głąbowski, D. Kmiecik, and M. Stasiak, "Overflow of elastic traffic," in *Proceedings of the 1st International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications, CoBCom 2016*, Austria, September 2016.

[62] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modelling overflow systems with distributed secondary resources," *Computer Networks*, vol. 108, pp. 171–183, 2016.

[63] M. Stasiak, "Blocking probability in a limited-availability group carrying mixture of different multichannel traffic streams," *Annales des Télécommunications*, vol. 48, no. 1-2, pp. 71–76, 1993.

[64] M. Głąbowski and M. Stasiak, "Multi-rate model of the group of separated transmission links of various capacities," in *Telecommunications and Networking—ICT 2004, 11th International Conference on Telecommunications, Fortaleza, Brazil, August 1–6, 2004, Proceedings*, J. N. de Souza, P. Dini, and P. Lorenz,

Eds., vol. 3124 of *Lecture Notes in Computer Science*, pp. 1101–1106, Springer, Berlin, Germany, 2004.

[65] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Analytical Modelling of Multi-tier Cellular Networks with Traffic Overflow," in *Computer Networks*, vol. 718 of *Communications in Computer and Information Science*, pp. 256–268, Springer International Publishing, Cham, Switzerland, 2017.

[66] ITU-T, "Calculation of the number of circuits in a group carrying overflow traffic," Recommendation E.521, 1988.

[67] T. Bonald and J. Roberts, "Internet and the Erlang formula," *ACM Computer Communications Review*, vol. 42, pp. 24–30, 2012.

[68] H. Akimaru and K. Kawashima, *Teletraffic: Theory and Application*, Springer, Berlin, Germany, 1999.

[69] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, no. 3-4, pp. 387–396, 1994.

[70] V. Paxson and S. Floyd, "Wide-area traffic: the failure of traffic modeling," in *Proceedings of the 1994 SIGCOMM Conference*, pp. 257–268, 1994.

[71] S. Floyd and V. Paxson, "Difficulties in simulating the Internet," *IEEE/ACM Transactions on Networking*, vol. 9, no. 4, pp. 392–403.

[72] J. Y. Hui, "Resource allocation in broadband networks," *Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1598–1608, 1988.

[73] F. Kelly, *Notes on Effective Bandwidth*, University of Cambridge, 1996.

[74] J. Roberts, *Performance Evaluation and Design of Multiservice Networks, Final Report COST 224*, Commission of the European Communities, Brussels, Belgium, 1992.

[75] J. Roberts, V. Mocci, and I. Virtamo, *Broadband Network Teletraffic, Final Report of Action COST 242*, Commission of the European Communities, Springer, Berlin, Germany, 1996.

[76] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968–981.

[77] I. Norros, "On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 953–962, 1995.

[78] A. Pras, L. Nieuwenhuis, R. van de Meent, and M. Mandjes, "Dimensioning network links: a new look at equivalent bandwidth," *IEEE Network*, vol. 23, no. 2, pp. 5–10, 2009.

[79] I. D. Moscholios, M. D. Logothetis, and A. C. Boucouvalas, "Blocking probabilities of elastic and adaptive calls in the Erlang multirate loss model under the threshold policy," *Telecommunication Systems*, vol. 62, no. 1, pp. 245–262, 2016.

[80] I. D. Moscholios, M. D. Logothetis, J. S. Vardakas, and A. C. Boucouvalas, "Congestion probabilities of elastic and adaptive calls in Erlang-Engset multirate loss models under the threshold and bandwidth reservation policies," *Computer Networks*, vol. 92, part 1, pp. 1–23, 2015.

[81] E. Brockmeyer, H. Halstrom, and A. Jensen, "The life and works of A.K. Erlang," *Acta Polytechnica Scandinavica*, vol. 6, no. 287, 1960.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 4163612, 13 pages
https://doi.org/10.1155/2018/4163612

Research Article

A Service-Oriented Approach for Radio Resource Management in Virtual RANs

Behnam Rouzbehani,[1] **Luis M. Correia**,[1] **and Luísa Caeiro**[2]

[1] *IST/INESC-ID, University of Lisbon, Lisbon, Portugal*
[2] *ESTG/INESC-ID, Setúbal Polytechnic Institute, Setúbal, Portugal*

Correspondence should be addressed to Behnam Rouzbehani; behnam.rouzbehani@tecnico.ulisboa.pt

Received 19 February 2018; Revised 25 May 2018; Accepted 12 June 2018; Published 10 July 2018

Academic Editor: Anna Zakrzewska

Virtualisation, as a key role player of future mobile communications, promotes the idea of service-oriented architectures. This paper proposes a model of Radio Resource Management (RRM) for emerging Virtual Radio Access Networks, based on the interaction between two separated management entities: Common-RRM (CRRM) to coordinate the radio resources among the Radio Access Technologies (RATs) and a centralised virtualisation platform on top of it, called Virtual-RRM (VRRM), which is responsible for service orchestration among Virtual Network Operators, enabling the definition of various services and policies, separately from vendors and underlying RATs. The main objective of VRRM is to satisfy the Service Level Agreements associated with different service classes to the highest possible level, within the framework of proportional fairness. On the other hand, CRRM is in charge of mapping the demanded capacity of each service onto the most suitable RATs. The model is further extended to deal with extreme situations of resource shortage, resulting from high traffic loads, by introducing delay to lower priority services. The performance of the proposed model is evaluated in a practical multi-RAT scenario. Results confirm that the isolation of service classes is consistent with the introduced serving weights, while all the demanded capacities from different services are responded by the most suitable RATs. Finally, independent of the variation of traffic load, 100% of the aggregated capacity is used.

1. Introduction and Motivation

The rapid increase of demand for new services in recent years has imposed crucial requirements to network operators. However, service provision in most of the existing telecom networks is tightly coupled with costly inflexible infrastructure, which has been specifically designed to deliver a limited range of services [1]. In order to overcome this barrier, management and operation functions have to be customised, to enable support for particular service types. Virtualisation, as a key enabler of the *Network-as-a-Service* paradigm for future mobile communications, can significantly facilitate the provision of customised services, by decoupling networking functionalities from the underlying infrastructure. In this framework, a common physical infrastructure can be effectively shared, in an isolated manner, among coexisting tenants, called Virtual Network Operators (VNOs); each VNO will be able to implement its own specific rules and regulations over the shared set of resources, without necessarily being aware of the underlying virtualisation process [2]. This is of great importance for the design of convergent mechanisms and coordination of the available resources across the emerging wireless Heterogeneous Networks (Het-Nets). In this regard, the main difference between the services that can be provided by future mobile implementations, compared to the majority of current technologies, such as LTE, lies in the *granularity* of the level of service customisation [3]. While in LTE all packets in a bearer are treated the same way, service flows in Virtual Radio Access Network (RAN) can be flexibly customised in a more granular way to sufficiently differentiate in between their requirements.

Radio Resource Management (RRM) is one of the key functionalities of cellular networks, which has a direct influence on the Quality of Service (QoS) of users, as well as on

the performance of higher layers [4]. With the introduction of new applications, which has led to the increased number of connected users, the problem of RRM has become particularly challenging, since the various applications have different and often conflicting needs [5]. Therefore, each service has to be managed independently, and the available radio resources must be allocated on a priority basis. This is the point when the context of *resource slicing* becomes interesting: each resource slice is defined to address the specific requirements of a service with a certain degree of performance isolation, to ensure that, regardless of the variation of network status, the desired performance level of independent slices is always met [6]. Such a flexible mechanism of radio resource slicing and management can be achieved through the virtualisation of radio resources. This way, an RRM algorithm should not only maximise the performance of different slices, but also the usage of the overall pool of shared resources [7].

Although there are quite extensive studies to address different challenges in traditional wireless networks, when it comes to the evolutionary technologies for future wireless communications, the existing techniques have to be modified in order to accommodate the specifications of new services and architectures. Specifically concerning the problem of RRM in Virtual RAN, there is a lack of effort to thoroughly cover the key parameters, such as *customised* service provisions, *isolation* between virtual slices, *fairness,* and the mechanisms of *interaction* between different entities. In [8], a model for RAN virtualisation is proposed, which provides a solution for the mapping of virtual network elements onto radio resources, as well as providing an algorithm of radio resource negotiation and allocation in a general term, which does not include specifications of different services into account. The problem of slice scheduling and performance isolation in Virtual RAN is addressed in [9], but, like the previous work, different needs of applications in network slicing, as well as fairness, are neglected. Although a fairness concept based on roaming price is defined in [10], and some QoS parameters are addressed in the proposed generalised virtualisation framework, the matter of isolation between services and operators is not covered.

Economic models for pricing based on optimising an objective function have been proposed with the purpose of balancing network throughput and users' fairness as two competing interests [11]. Among the available mechanisms, proportional fairness has proved to be an effective approach when the objective is to maximise the average long-term users' data rate [12, 13]. In this regard, it is suggested that, by employing a logarithmic utility function in the slice scheduler of Virtual RAN, an effective mechanism of fairness based on the concept of proportional fairness can be achieved among virtual slices [14]. A model of RRM is proposed in [15] for 5G wireless infrastructures, which aggregates users' traffic across a Wi-Fi/LTE network, by considering an α-fair mechanism in the objective function, including proportional fairness as its special case; however, there is no effort to address the specifications of different services. Another approach for cooperative RRM in 5G heterogeneous cloud RANs is proposed in [16], which considers a modified max-min technique as an alternative to proportional fairness, although still the evaluation of the effect of different service parameters on model performance is neglected.

This paper proposes a model for *RAN-as-a-Service*, which deals with the *high-level* management of Virtual RAN, considering the concept of network slicing according to the specific customised service requirements. The contribution of this work is twofold. First, the proposed models of Virtual-RRM (VRRM) in previous works [17, 18] are extended and modified in order to realise the *separation* in the role of Infrastructure Providers (InP) and VNOs. This way, the task of mapping the demanded capacity from different virtual slices onto the underlying physical RATs is defined in Common-RRM (CRRM), in order to promote the notion of *end-to-end slicing*, since it was not covered previously. To accommodate this function, a new mechanism of *cooperation* between the two entities (i.e., VRRM and CRRM) is also proposed to define which information has to be exchanged in order to achieve an efficient interaction, while keeping a desired level of isolation.

Second, regarding the fairness problem, while in the previous paper [17], the definition of fairness is to minimise the deviation of services data rates from a nominal fixed value, the framework has been changed to a more flexible and accurate one, to cope with the concept of *proportional fairness.* This change comes with the price of a more complex objective function compared to the previous linear one. However, since VRRM deals with the high-level network management, including VNOs' policies, which does not require to be changed so often, this level of complexity is tolerable, while the mechanism of CRRM is still a low complexity linear mapping from the capacity demands onto the underlying suitable RATs.

It is worth mentioning that although some user-based metrics, such as users' average data rate or average percentage of delayed users, are defined to evaluate the model, the current work is generally not intended to address the traditional problem of RRM at the *end-user scale* or the associated metrics, such as service delay time. The reason is that the slice scheduling of service flows is performed in upper layers, and also that this scheduling does not need to further modify the existing mechanisms of RRU schedulers in lower layers [19]. Furthermore, since this work deals with a higher layer management of the network, the objectives can be addressed by an abstract vision of the underlying infrastructure. Accordingly, more detailed assumptions for the lower layers, such as power control, antenna configuration, scheduling, mobility management, channel information, and protocols, are out of the scope of this paper. Regarding the *scalability* of VRRM with a proportional fairness mechanism of scheduling, as well as service orchestration in *multitenant* environment, another work has been already developed [20], which specifically addresses these issues and proves the efficiency of the proposed VRRM model; therefore, neither of these topics are covered in this work.

The rest of the paper is organised as follows. Section 2 represents the conceptual network architecture along with the functionalities of the involved components. Section 3 describes the analytical model of VRRM, CRRM, and the mechanism of interaction in between each other. In order to

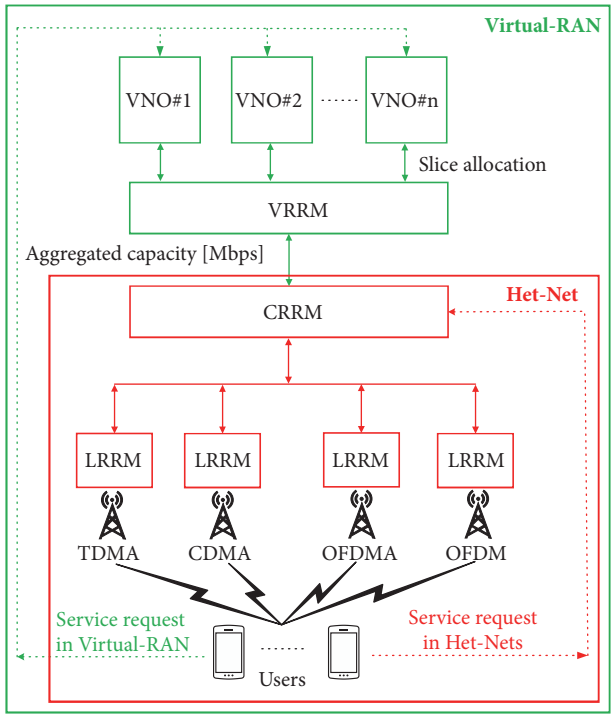


Figure 1: Differences between traditional Het-Net and Virtual RAN.

evaluate the proposed model, a reference scenario with some assumptions is defined in Section 4, followed by the analysis of results obtained from simulations in Section 5. Finally, the paper is concluded in Section 6, by highlighting the main achievements.

2. Network Architecture and Main Assumptions

In order to point out the main differences between traditional Het-Nets and the proposed model for Virtual RANs and to explain the functionalities of the involved parties in the RRM model, a software-based hierarchical network architecture is shown in Figure 1, which is a conceptual architecture consistent with the suggested 3GPP business model for *wholesale-only* network sharing, in which Infrastructure Providers (InPs) do not offer service to end users, rather selling capacity to businesses that do not own the infrastructure [14] (details on physical implementations as well as on different interface protocol features can be found in [21]).

The service connectivity request in typical Het-Nets is directed to CRRM, as the usual entity for network management, and processed centrally to be assigned to a suitable RAT according to a decision criterion. However, in the proposed architecture for Virtual RAN, demand for a specific service and capacity goes directly to the linked VNO as the service provider. This capacity needs to be delivered respecting the Service Level Agreements (SLAs) between InP and VNO.

In contrast to the existing Het-Nets, the role of network operators is separated from InPs; accordingly, VNOs on top of the hierarchy do not own the infrastructure, rather sharing the radio resources from different RATs owned by InPs. As a result, from their perspective, it is not important by which technology they are being served, as long as the SLAs are satisfied. VNOs ask for *Capacity-as-a-Service* from a centralised virtualisation platform called VRRM [17], which does not exist in the current architecture of Het-Nets and is placed on the top of CRRM. VRRM is in charge of managing the total available capacity provided by CRRM, through aggregating all the Radio Resource Units (RRUs) from different RATs, which can be OFDM (related to Wi-Fi), OFDMA (related to LTE), CDMA (related to UMTS), and TDMA (related to GSM), and sharing the capacity from separated slices associated with different services of VNOs. By providing isolation and element abstraction, VRRM enables each VNO to deploy its own protocol stack over the same set of RRUs per RAT (e.g., resource-blocks in LTE, codes in UMTS, time-slots in GSM, and carriers in Wi-Fi) [22], therefore, promoting the notion of multitenancy in a virtualised environment with several existing access techniques.

VRRM has to closely interact with CRRM, by translating VNOs' requirements and different SLAs into a set of management policies for the lower levels [17]. These policies contain information about the capacity demanded from each service, as well as their priorities. In return, CRRM provides VRRM with the monitoring reports and information (e.g., the available aggregated capacity) to enhance its performance. CRRM is in charge of mapping the resources of different RATs to satisfy the requests of VRRM, by demanding each RAT to provide a portion of the available capacity to be assigned to end users. The service-to-RAT association mechanism takes the load and suitability of each RAT for performing specific services into account. The performance of CRRM is being optimised, based on the information coming from Local-RRMs (LRRMs), which are in charge of managing the allocation of local RRUs from each RAT to the connected end users.

It is assumed that VNOs can provide the four service classes [23], i.e., *conversational*, *streaming*, *interactive,* and *background*, considering that the design and implementation of future mobile communications, including 5G RAN architecture, should support the 3GPP service classes [4]. As mentioned, since service customisation is one of the main aspects of RAN slicing, on top of these service classes three types of SLAs are also proposed in this work to define the level of service guarantees in terms of priority and contracted capacity, which can be modified according to the VNOs' policies. The three general categories of SLAs are considered as follows:

(i) Guaranteed bitrate (GB): the highest priority category, for which minimum and maximum thresholds for data rate assignment have to be always guaranteed, regardless of the variation of traffic load and network status; therefore, users are always expecting a good quality in return of a relatively higher service price.

(ii) Best effort with minimum guaranteed (BG): the second highest priority, for which just a minimum level of data rate is guaranteed, and higher data rates are served in a best effort manner in case of availability;

from the users' viewpoint, a service with acceptable quality and affordable price is expected.

(iii) Best effort (BE): the lowest priority type, for which there is no level of service guarantees and users are served in a pure best effort manner; consequently, in extreme case of high traffic loads, BE users are the first ones who suffer.

These definitions of SLAs are in line with 5G network design assumptions, since the QoS profile of service flows should carry the information about whether a service traffic is categorised as GB or non-GB [3].

In addition, one should note that, although only data rates are being considered in here for the definition of the SLAs, in a more global perspective (which is out of the scope of this paper but is being considered for future work), one should take other parameters, like link latency, connection reliability, and connectivity capacity.

3. Radio Resource Allocation Model

3.1. VRRM Approach. The primary goal of VRRM is to maximise the usage of the aggregated capacity, which is calculated and provided by CRRM, in order to satisfy the contracted SLAs to the highest possible level, considering services' priority, while distributing capacity according to the concept of proportional fairness [24], subject to some constraints, including maximum achievable capacity, predefined SLA thresholds, and access of users to different RATs.

To realise these goals, VRRM's analytical model is formulated as a constrained concave optimisation problem. The objective function in (1) is defined with the aim of balancing between efficiency and fairness when allocating the resources in a network with heterogeneous services. The utility function, f_{VRRM}, is a measure of efficiency, mapping the portion of network bandwidth assigned to users onto a real number, quantifying the expected users' satisfaction, given the allocated resources; the logarithmic behaviour implies that users' satisfaction increases with the increasing rate as the allocated bandwidth increases. f_{VRRM} is executed in independent time intervals, representing the VRRM decision windows according to the framework of [25]. In this regard, all network parameters, such as number of users or allocated data rates, are assumed to be fixed during each time interval, representing the parameter's average value in this period.

$$\max_{\boldsymbol{w}^{usr}} f_{VRRM}\left(\mathbf{w}^{\mathbf{usr}}\right) = \max_{\mathbf{w}^{\mathbf{usr}}} \sum_{k=1}^{N^{srv}} \lambda_k \log\left(\sum_{i=1}^{N_k^{usr}} w_{k,i}^{usr} \frac{R_{k[\mathrm{Mbps}]}^{srv_{max}}}{R_{[\mathrm{Mbps}]}^{CRRM}}\right) \tag{1}$$

where

(i) N_k^{usr}: number of users performing service k,

(ii) N^{srv}: number of provided services,

(iii) $R_k^{srv_{max}}$: maximum assignable data rate for service k,

(iv) R^{CRRM}: total available capacity provided by CRRM to VRRM,

(v) $w_{k,i}^{usr}$: assigned weight to user i, performing service k, ranging in $[0, 1]$,

(vi) λ_k: weight of service k, prioritising data rate assignment,

(vii) $\mathbf{w}^{\mathbf{usr}}$: vector of users' weights, to obtain the long-term average users' data rate, according to the contracted SLAs provided by VNOs to VRRM, as well as other constraints:

$$\mathbf{w}^{\mathbf{usr}} = \left[w_{1,1}^{usr}, w_{1,2}^{usr}, \ldots, w_{1,N_1^{usr}}^{usr}, \ldots, w_{N^{srv},1}^{usr}, w_{N^{srv},2}^{usr}, \ldots, w_{N^{srv},N_k^{usr}}^{usr}\right]^T \tag{2}$$

This way, VRRM builds a bridge between the functionalities of MAC and higher layers, by optimising the allocation of radio resources for different applications [26]. Serving weights λ_k define services' priorities and enable the tuning of the portion of the data rate that will be assigned to each service, being assumed that they are positive integer numbers: a service with higher priority has a higher serving weight, and the lowest value is always assumed to be 1. On the other hand, user's weights, $w_{k,i}^{usr}$, are the desired parameters to be obtained by solving the optimisation problem, indicating the share of each user from the maximum achievable data rate. Therefore, the average long-term data rate of each user has to fall in the acceptable range of data rate variation defined in the SLA contracts as follows:

$$R_{k[\mathrm{Mbps}]}^{srv_{min}} \leq w_{k,i}^{usr} R_{k[\mathrm{Mbps}]}^{srv_{max}} \leq R_{k[\mathrm{Mbps}]}^{srv_{max}} \tag{3}$$

where

(i) $R_k^{srv_{min}}$: minimum assignable data rate for service k.

In addition, there is a logical constraint, which specifies that the total bandwidth allocated to all users performing different services cannot surpass the total aggregated capacity provided by CRRM. Therefore, the whole network capacity is subject to an upper bound for the total used bandwidth assigned to users:

$$\sum_{k=1}^{N^{srv}} \sum_{i=1}^{N_k^{usr}} w_{k,i}^{usr} R_{k[\mathrm{Mbps}]}^{srv_{max}} \leq R_{[\mathrm{Mbps}]}^{CRRM} \tag{4}$$

The last constraint indicates the fact that the access of users to different RATs depends on their location. Assuming that a certain number of users are enjoying full coverage of all available RATs, while the rest has access just to the cellular ones, the following condition can be used to define the limit of access of the latter group to cellular RATs:

$$\sum_{k=1}^{N^{srv}} \sum_{i=1}^{N_k^{usr_{cell}}} w_{k,i}^{usr} R_{k[\mathrm{Mbps}]}^{srv_{max}} \leq R_{[\mathrm{Mbps}]}^{CRRM_{cell}} \tag{5}$$

where

(i) $N_k^{usr_{cell}}$: number of users with access only to cellular RATs, performing service k,

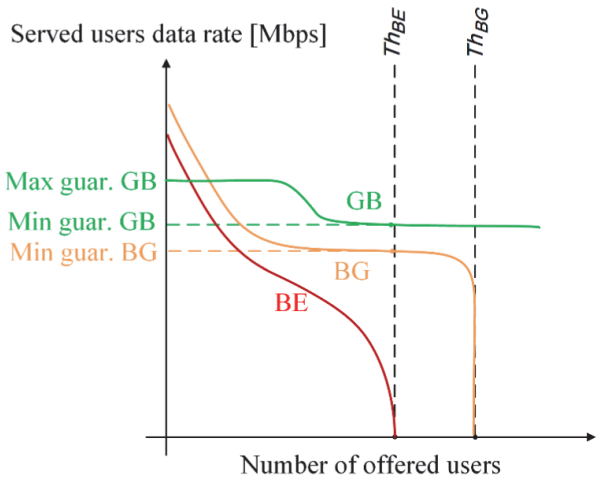


Figure 2: General policy of data rate allocation for different services.

(ii) $R^{CRRM_{cell}}$: aggregated capacity obtained from the RRUs of cellular RATs, provided by CRRM.

The proposed problem is solved by CVX [27], which is a modelling system based on MATLAB, developed by Stanford University for disciplined convex programming. The method used to solve the problem is primal-dual interior point [28], the termination tolerance of the function being set to 10^{-5}.

The algorithms of slicing and data rate allocation are based on *prioritisation* of services and network pricing, which also include an admission control policy to guarantee the required SLAs, being a common approach for service-based RAN slicing [29, 30]. Moreover, in the extreme situation when network capacity is not enough to satisfy all users with at least the minimum guaranteed service level, delay is introduced to some of the low priority users, releasing the required capacity to serve the remaining ones with the minimum acceptable data rate. In this respect, the number of delayed users during each VRRM decision window corresponds to the average number of users whose service requests are not granted during this period. The general trend for data rate allocation is illustrated in Figure 2, representing the expected behaviour of VRRM under the specific model assumptions. For an increasing number of users, up to threshold Th_{BE} all users are served, but after that, BE users are discarded, since they have the lowest priority, in order to provide VRRM the possibility of serving all BG and GB ones with the minimum contracted data rates; after this point, capacity is not sufficient to address all contracted data rates; therefore, a reasonable decision is to start delaying just enough number of BG users to free capacity for GB ones. Threshold Th_{BG} represents the point when no BG user is left to be delayed, and the VRRM mechanism has to start delaying some of the GB ones as the last alternative.

3.2. Interaction between CRRM and VRRM. Since VRRM is a centralised entity, responsible for allocating the required capacity of each service defined by the associated VNO, based on the specific policies of that VNO regarding the contracted SLAs and serving priorities, CRRM has to provide VRRM with the information about the maximum aggregated capacity, which is accessible from all available RATs. Then, based on this information, VRRM calculates the optimum way of distributing capacity among the different services, to satisfy their requirements. One should note that CRRM does not pass the information about the available capacity of individual RATs to VRRM, since VNOs do not own the physical infrastructure, and VRRM is not concerned about dealing with the management of the resources of the various RATs.

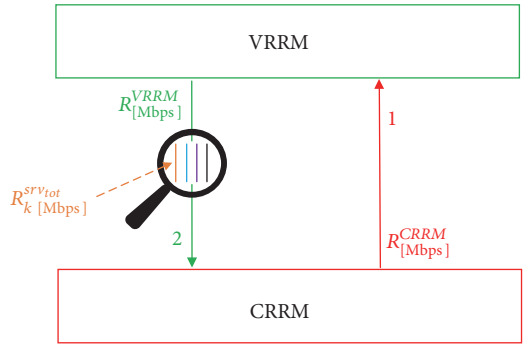


Figure 3: Interaction between CRRM and VRRM.

On the other hand, the total demanded data rate of each service, which is calculated by VRRM, should be assigned to suitable RATs, according to the specification of each access technology and requirements of different services. This way, VRRM has also influence on the decision procedure of CRRM. It is also notable that VRRM does not pass the information about the policies of individual VNOs for serving their users, such as contracted SLAs, to CRRM. Figure 3 represents the level of interaction between CRRM and VRRM.

Information about the available aggregated capacity, R^{CRRM}, takes place in the first step, provided by CRRM as a VRRM decision-making parameter. Then in the next step, VRRM calculates the total demanded capacity of each service, $R_k^{srv_{tot}}$, to be returned to CRRM for RAT selection process; R^{VRRM} represents the aggregation of all the calculated demands performed by VRRM.

3.3. Service-to-RAT Assignment Mechanism of CRRM. In order to take advantage of the specifications of different technologies in a multi RAT network and to achieve multiplexing gain, it is necessary to define a precise cooperation mechanism among RATs. The proposed technique used for mapping demand onto the available capacity of each RAT is fundamentally a *network-centric* approach, since the main objective is to maximise the *global* utilisation of radio resources. However, since the contracted SLAs are considered as a decision-making parameter in the calculation of demands by VRRM, a preestablished user's satisfaction level is also projected as a constraint to be satisfied by the CRRM model. A *service-based* policy is applied to the process, in order to ensure that mapping the various services onto the

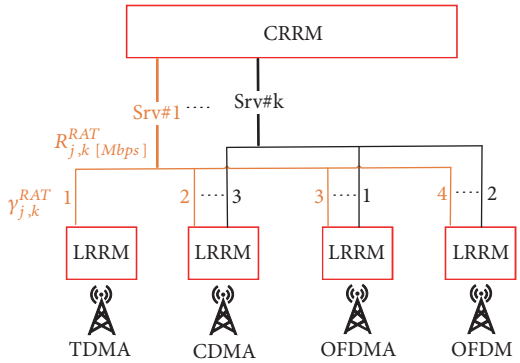


Figure 4: CRRM mechanism of mapping demands to the most suitable RATs.

different access technologies is based on the feasibility and suitability of both characteristics [31]; e.g., it is not feasible to serve a video streaming user with GSM, or it is preferable to use Wi-Fi for file sharing. Accordingly, for a specific service type, a prioritised list of suitable RATs is defined in the model to arrange the available common radio resources [32].

To address the assumptions regarding the maximisation of the total utilisation of radio resources, as well as the distribution of the traffic load among existing RATs by considering a service-based policy, the objective of CRRM can be expressed as a weighted linear function, f_{CRRM}, which maps a portion of the available capacity from each RAT onto the most suitable service:

$$\max_{\mathbf{R^{RAT}}} f_{CRRM}\left(\mathbf{R^{RAT}}\right) = \max_{\mathbf{R^{RAT}}} \sum_{j=1}^{N^{RAT}} \sum_{k=1}^{N^{srv}} \gamma^{RAT}_{j,k} \frac{R^{RAT}_{j,k[\text{Mbps}]}}{R^{CRRM}_{[\text{Mbps}]}} \tag{6}$$

where

(i) N^{RAT}: number of RATs,

(ii) $\gamma^{RAT}_{j,k}$: assigned weight to RAT j, for performing service k, to define service-to-RAT allocation priorities,

(iii) $R^{RAT}_{j,k}$: assigned data rate from RAT j to service k,

(iv) $\mathbf{R^{RAT}}$: vector of assigned data rates:

$$\mathbf{R^{RAT}} = \left[R^{RAT}_{1,1}, R^{RAT}_{1,2}, \ldots, R^{RAT}_{1,N^{srv}}, \ldots, R^{RAT}_{N^{RAT},1}, R^{RAT}_{N^{RAT},2}, \ldots, R^{RAT}_{N^{RAT},N^{srv}}\right]^T \tag{7}$$

Figure 4 represents the mechanism of f_{CRRM} in distributing the traffic load associated with the different services, among the existing RATs. When a service demand is made to CRRM, the traffic load is distributed among the available RATs, with a higher portion of the load directed to the most suitable ones, in case there is enough capacity from those RATs to serve the demand. Accordingly, the services with higher priority have the possibility of using the most suitable RATs, while the remaining demand from low priority services is more likely to be directed to less suitable ones, especially under an extreme case of high traffic load.

Furthermore, there are some constraints that should be taken into consideration while optimising the objective function. An important constraint is the one that projects the demanded data rate of each service from VRRM and plays an important role in the interaction between CRRM and VRRM. From Figure 4, it is noticeable that the distribution of traffic load for each service among the different RATs should sum to the demanded data rate of that specific traffic slice, provided by VRRM; it should not be either higher, since the extra capacity will be wasted, or lower, because contracted SLAs cannot be addressed. This constraint can be expressed as follows:

$$\sum_{j=1}^{N^{RAT}} R^{RAT}_{j,k[\text{Mbps}]} = R^{srv_{tot}}_{k[\text{Mbps}]} \tag{8}$$

The next constraint is related to the capacity of each RAT. The aggregation of demands that are directed to each RAT cannot exceed the total available capacity of that RAT, which can be written as

$$\sum_{k=1}^{N^{srv}} R^{RAT}_{j,k[\text{Mbps}]} \leq R^{RAT_{tot}}_{j[\text{Mbps}]} \tag{9}$$

where

(i) $R^{RAT_{tot}}_{j}$: total available capacity of RAT j.

The last constraint ensures that all directed loads from services to RATs are positive values:

$$R^{RAT}_{j,k[\text{Mbps}]} \geq 0 \tag{10}$$

The simplex method is adopted to find the optimal problem solution. Instead of evaluating all the possible candidates in the feasible solution space that satisfy constraints, this method examines just the *better* candidates, which are known in advance, yielding larger values of the objective function [33].

3.4. Calculating the Aggregated Capacity. The achievable data rate of each RRU depends on the SINR level of its associated user. Based on the model suggested in [34], considering an interference limited Het-Net, with the assumption that users experience channel fading with a Rayleigh Distribution (hence, the received power being described by the Exponentially Distribution), the probability of having SINR higher than an arbitrary value is derived. The authors in [17] further modified the model to find the Probability Distribution Function (PDF) of a single RRU's data rate, assuming that it varies between minimum and maximum positive values:

$$P_{\text{R[Mbps]}}\left(R^{RRU}_{j} \mid 0 \leq R^{RRU_L}_{j} \leq R^{RRU}_{j} \leq R^{RRU_H}_{j}\right) = \frac{e^{-(0.2/\alpha_p)\ln(10)\sum_{m=0}^{5} a_m (R^{RRU_L}_{j[\text{Mbps}]})^m} - e^{-(0.2/\alpha_p)\ln(10)\sum_{m=0}^{5} a_m (R^{RRU}_{j[\text{Mbps}]})^m}}{e^{-(0.2/\alpha_p)\ln(10)\sum_{m=0}^{5} a_m (R^{RRU_L}_{j[\text{Mbps}]})^m} - e^{-(0.2/\alpha_p)\ln(10)\sum_{m=0}^{5} a_m (R^{RRU_H}_{j[\text{Mbps}]})^m}} \tag{11}$$

where

(i) $R^{RRU_L}_{j}$: lower bound for the RRU's data rate,

Table 1: Factor of polynomial approximation (updated from [35]).

Coefficients	TDMA (GSM)	CDMA (UMTS)	OFDMA (LTE)	OFDM (Wi-Fi)
a_0	7.95	-12.15	0.24	1.38
a_1	964.6	1.89	264.1	189.3
a_2	-293.8	-0.041	34.96	29.76
a_3	0	0	-274.4	-351.8
a_4	-871.2	0	-52.77	-38.91
a_5	109.9	0	309.8	411.5

(ii) $R_j^{RRU_H}$: higher bound for the RRU's data rate,

(iii) α_p: path loss exponent, where $\alpha_p \geq 2$,

(iv) a_m: coefficients of a 5-degree polynomial, based on real network logs, Table 1.

The goal of deriving the PDF of data rate for each RRU is to find the total capacity of each RAT, as well as the aggregated network capacity, being provided from CRRM to VRRM. Therefore, the next step is to obtain the distribution functions associated with the total capacity of each existing RAT. Assuming that RAT j can assign a N_j^{RRU} number of RRUs to the connected users and that all channels are independent, the data rates of all RRUs, which are random variables, are independent as well. Therefore, the PDF of the accumulated data rate of each access technology can be expressed as the convolution of all its RRU's PDFs [36]:

$$p_{\mathrm{R[Mbps]}}\left(R_j^{RAT_{tot}}\right) = p_{\mathrm{R[Mbps]}}\left(R_j^{RRU_1}\right) * p_{\mathrm{R[Mbps]}}\left(R_j^{RRU_2}\right) * \cdots * p_{\mathrm{R[Mbps]}}\left(R_j^{RRU_{N_j^{RRU}}}\right) \quad (12)$$

According to the Central Limit Theorem, the convolution PDF of statistically independent random variables tends to a Normal Distribution, as the number of random variables tends to infinity, regardless of the initial PDFs [36]. Since the random variables in this work are RRUs, which are proposed to vary between a minimum and a maximum value for all the RATs, (12) is approximated by fitting a Truncated Normal Distribution for each RAT [37]. Assuming that $R_j^{RAT_{tot}}$ has a Normal Distribution, $R_j^{RAT_{tot}} \sim N(\overline{R_{b_j}}, \sigma_j)$, and that the interval for R_j^{RRU} is $R_j^{RRU} \epsilon [R_j^{RRU_H}/2, R_j^{RRU_H}]$, then, the distribution parameters associated with each RAT within 95% confidence bounds are given in Table 2, where $R_j^{RAT_{tot}} \epsilon [R_{b_j}^{min}, R_{b_j}^{max}]$.

The general steps taken to obtain the convolution PDF are summarised in what follows. First, the relationship between SINR and the data rate of a single RRU is obtained. The information for UMTS and LTE is based on real measurements provided by Portuguese mobile operators [38], the rest being obtained according to the theoretical behaviour of SINR as a function of data rate. Considering that data rate is constant in given SINR intervals, SINR can be represented as a piecewise step function of data rate and further estimated by fitting a continuous fifth-degree polynomial, using the least-square technique. Then, following (11), the PDF of a single RRU is calculated, and finally the aggregation of PDFs in the convolution function of each RAT (12) yields the result.

Table 2: Parameters of the truncated normal distributions.

RATs	$R_{b_j[\mathrm{Mbps}]}^{\min}$	$R_{b_j[\mathrm{Mbps}]}^{\max}$	$\overline{R_{b_j[\mathrm{Mbps}]}}$	$\sigma_{j[\mathrm{Mbps}]}$
OFDM (Wi-Fi)	3352	6704	5116	67.1
OFDMA (LTE)	2400	4800	3655	61.2
CDMA (UMTS)	44.1	88.2	66.2	1.63
TDMA (GSM)	0.62	1.24	0.94	0.053

4. Case Study Scenario

Model performance is evaluated by defining a case study scenario. It is assumed that the area under analysis is 1 km^2, which is uniformly covered by all cellular RATs as shown in Figure 5(a). Furthermore, a Wi-Fi Access Point (AP) based on IEEE802.11ac is placed at the centre of each LTE cell site to boost capacity. These assumptions of cell layout are just practical examples that are common among the studies that discuss the problem of resource management in 5G or future implementations from a flow level perspective [39, 40]. In this regard, the aim of considering a scenario with coexisting different RATs is to evaluate the performance of CRRM in terms of mapping service demands onto RATs according to the suitability and loading factor of these RATs. Concerning the distribution of users as presented in Figure 5(b), 25% are gathered around the *centre (ce)* of LTE Base Stations (BSs), while the remaining 75% are located in *off-centre (oc)* areas, without access to Wi-Fi, but with full access to cellular RATs. Each user performs one specific service at a time.

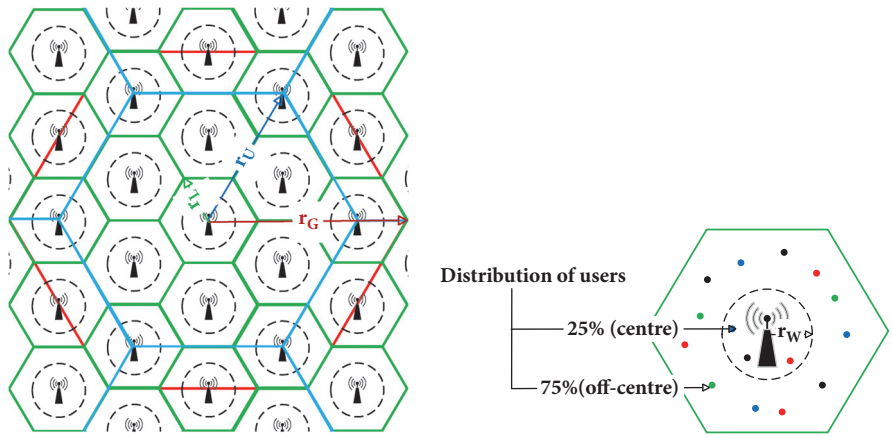
Specifications are summarised in Tables 3 and 4, where r_j^{Cell} is the coverage radius of a single BS or AP, $R_j^{RRU_{\max}}$ being the maximum data rate of each RRU from different RATs. However, these values are not achievable in practice, as the quality of the physical channel is greatly influenced by SINR. Using the proposed convolution PDFs to calculate the aggregated capacity of each RAT, by choosing $\alpha_p = 3.8$ as a common value for urban outdoor environments [17], and then randomly selecting from the associated convolution PDF of RATs, $R_j^{RAT_{tot}}$ as the average value for the total available capacity of each access technology *per km^2* can be obtained.

Table 3: RAT specifications.

RAT	# BS, AP	r_j^{Cell}	$r_{j[\text{km}]}^{Cell}$	RRU	$R_{j[\text{Mbps}]}^{RRU_{max}}$	$R_{j[\text{Mbps}]}^{RAT_{tot}}$
OFDM (Wi-Fi)	16	r_W	0.05	Carrier	0.97	504
OFDMA (LTE)	16	r_L	0.4	Res. block	0.75	350.2
CDMA (UMTS)	~1.7	r_U	1.2	Code	1.4	6.54
TDMA (GSM)	1	r_G	1.6	Time-slot	0.059	0.10

Table 4: Service parameters.

Service	Range of data rate [Mbps]		User mix [%]	λ_k	SLA
	Centre	Off-centre			
Voice (Vo) ***Conversational***	[0.032, 0.064]		20	50	GB
Video streaming (Vi) ***Streaming***	[2, 13]		50	30	GB
Web browsing (We) ***Interactive***	[1, 862.4]	[1, 358.4]	20	6	BG
Email (Em) ***Background***	[0, 862.4]	[0, 358.4]	10	1	BE



(a) Coverage of existing RATs (b) Distribution of users

Figure 5: Network layout for the reference scenario (r_G=1.6 km, r_U=1.2 km, r_L=0.4 km, and r_W=0.05 km).

Summing up over all these values, one gets R^{CRRM}, as 860.8 Mbps. In order to find the maximum achievable capacity that can be provided to *oc* users, Wi-Fi should be excluded from the process, leading to 358.4 Mbps for $R^{CRRM_{Cell}}$.

Concerning the assumptions for service parameters and SLAs, one service from each class is defined to be served by a single VNO. One can see in Table 4 that *Voice (Vo)* and *Video streaming (Vi)* are categorised as GB services, since they are delay sensitive and demanding almost constant data rates. Therefore, allocating capacities higher than the contracted ones will not improve their QoS [41]. On the other hand, an increase in the assigned data rate of *Web browsing (We)* and *Email (Em)* can indeed improve the users' quality of experience. The choice of these four services is according to the fact that up to 2023 more than 80% of the mobile traffic will be comprised of these services [42]. Also, concerning the assumptions for allocating the most suitable RATs to different services, the prioritised table of RAT selection for the proposed services is presented in Table 5. For each service, the RAT that is numbered as 1 is the most preferred access technology to be associated with that service, NA representing the case that is not feasible to associate a particular service to a specific RAT.

5. Analysis of Results

The numerical trend of data rate assignment, for *ce* and *oc* users, is shown in Figure 6. It is clear that, for the same service, the data rate allocated to *ce* users is always higher than or equal to the one of *oc* users, since Wi-Fi coverage is only available in *ce* areas. When the number of users is comparatively low, all GB users are well served; as the traffic load increases, the data rates of *oc* users drop to the lowest contracted level (defined in Table 4).

Th_{BE} is the point where the algorithm stops serving the BE users in order to be able to continue serving the higher priority ones with the minimum guaranteed data rates. It is also noticeable that *Vo* is the last service to drop from 64 kbps to 32 kbps just before Th_{BE}, since it has the highest priority and the lowest demanded data rate compared to the other services.

Table 5: Prioritised table of service to RAT assignment.

Services	Priority of RATs			
	TDMA (GSM)	CDMA (UMTS)	OFDMA (LTE)	OFDM (Wi-Fi)
Voice	1	2	3	4
Video streaming	NA	3	1	2
Web Browsing	4	3	2	1
Email	4	3	2	1

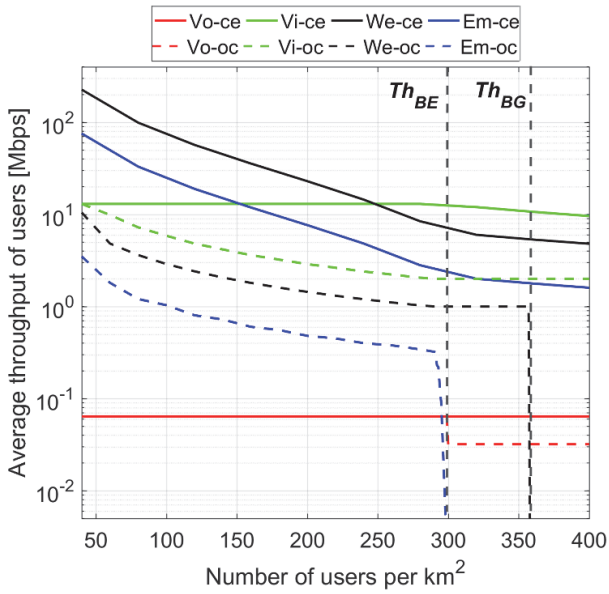


Figure 6: Average long-term data rate of users.

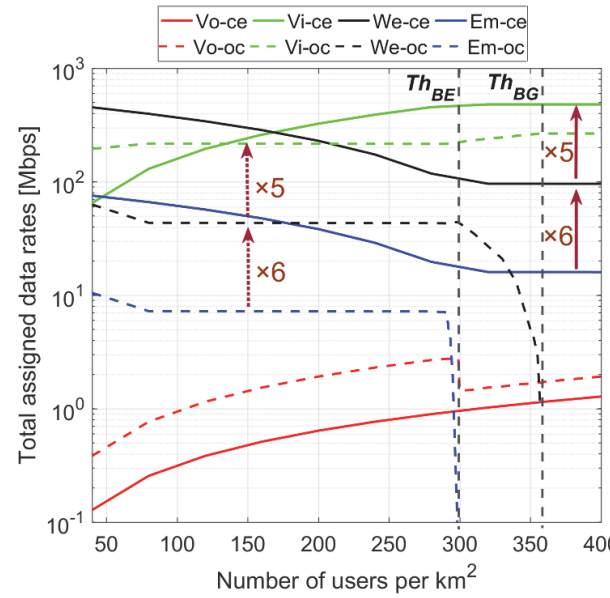


Figure 7: Total allocated capacity to the services.

Figure 7 shows the total data rate assigned to each service, as another evaluation metric. The results from this figure, including the different thresholds, are compatible with those of Figure 6. When there is no limitation for data rate assignment in both *ce* and *oc* areas, the ratio of the total data rate allocated to *Vi*, *We,* and *Em* is proportional to their service weights, 30, 6, and 1, respectively. The total used network capacity, one of the most important parameters, is the aggregation of data rate values at each point for both *ce* and *oc* users, being almost equal to the average total available capacity of VRRM obtained per km^2. Hence, it is shown that, independent of the variation of traffic load and configuration of several parameters in the proposed scenario, the VRRM algorithm is capable of distributing all the available capacity to address the demanded data rates based on the proposed SLAs.

The percentage of served users is shown in Figure 8. As all users are served before Th_{BE}, the values for both *ce* and *oc* in this part remain constant, being associated with the ones defined for the user mix in Table 4. The percentage of served *ce* users in each service does not change by increasing the number of offered users up to 400, as capacity is enough to serve them all. However, at Th_{BE}, the values for *Em-oc* go to zero, meaning that all BE users located in *oc* are delayed in order to provide enough capacity to serve the rest of *oc* users, leading to a slight increase in the values of other services.

After Th_{BE}, the algorithm starts discarding *We-oc*, as BG users. At this point, since a minimum data rate is guaranteed for BG users, the algorithm delays just enough number of *We-oc* ones to release capacity for the rest of *oc* users to continue their service with the minimum contracted data rates, in order to maximise serving the higher priority users up to Th_{BG}, where no BE users left to be delayed. The same process takes place after Th_{BG} for GB services, starting from *Vi-oc*. It is also noticeable that at Th_{BG}, all available capacity at *oc* is divided between *Vi* and *Vo*, the number of *Vi* users being 2.5 times higher than *Vo* ones, which is proportional to their ratio of traffic share (50% and 20% for *Vi* and *Vo*, respectively).

In order to evaluate the performance of the CRRM mechanism of service-to-RAT assignment, as well as the interaction between the CRRM and VRRM, the procedure of allocating capacity from the available RATs, to satisfy the demanded data rate of each service, is shown in Figure 9.

Starting from *Vo* in Figure 9(a), one can see that the demanded data rate is provided by GSM and UMTS in both *ce* and *oc* areas. As GSM is more suitable to serve voice users, the CRRM algorithm gives the higher priority to this RAT. However, the demanded data rate of voice is higher than the total available capacity of GSM; therefore, the rest of *Vo* traffic load is served by UMTS, which is the second highest priority RAT. One can also notice that before Th_{BE}, since there is no

Table 6: RATs and services matching, before Th_{BG}.

RAT	Voice		Video		Web		Email	
	Ce	Oc	Ce	Oc	Ce	Oc	Ce	Oc
TDMA (GSM)	√	√	NA	NA	×	×	×	×
CDMA (UMTS)	√	√	×	×	×	√	√	√
OFDMA (LTE)	×	×	√	√	×	√	×	√
OFDM (Wi-Fi)	×	NA	√	NA	√	NA	√	NA

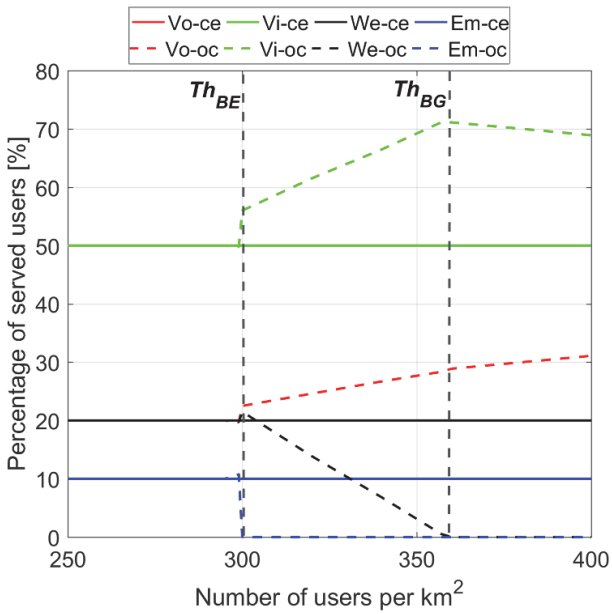


Figure 8: Percentage of served users.

limitation in data rate assignment, capacity share of *oc* users for both RATs is three times higher than the *ce* ones, which is in accordance with the ratio of traffic (i.e., 75% and 25% accordingly) in these two areas.

Regarding the *Vi* demand in Figure 9(b), although the number of *oc* users is three times higher than *ce* ones, the demanded data rate in centre is mostly higher, except for lower traffic rates, since this group of users have access to Wi-Fi. Accordingly, for *oc* users, the available LTE capacity as the most preferred RAT is enough to serve the demanded data rates until Th_{BE}. After that, *oc* users have to start using capacity from the second available preferred RAT, i.e., UMTS, while at each point for *ce* users the capacity of LTE and Wi-Fi is sufficient for the total demand.

For *We* users, the allocated capacity from different RATs is presented in Figure 9(c). It is noticeable that the whole traffic of *ce* users can be served just by Wi-Fi as the most preferred RAT. Regarding the *oc* users, before Th_{BE}, LTE as the most suitable RAT can cover the demand. However, by increasing the required data rate of *Vi-oc* users between Th_{BE} and Th_{BG}, there will not be enough capacity from LTE to serve the rest of *oc* traffic; therefore, CRRM will choose the second suitable RAT, UMTS, besides LTE to address all the demanded capacity of *We-oc* users.

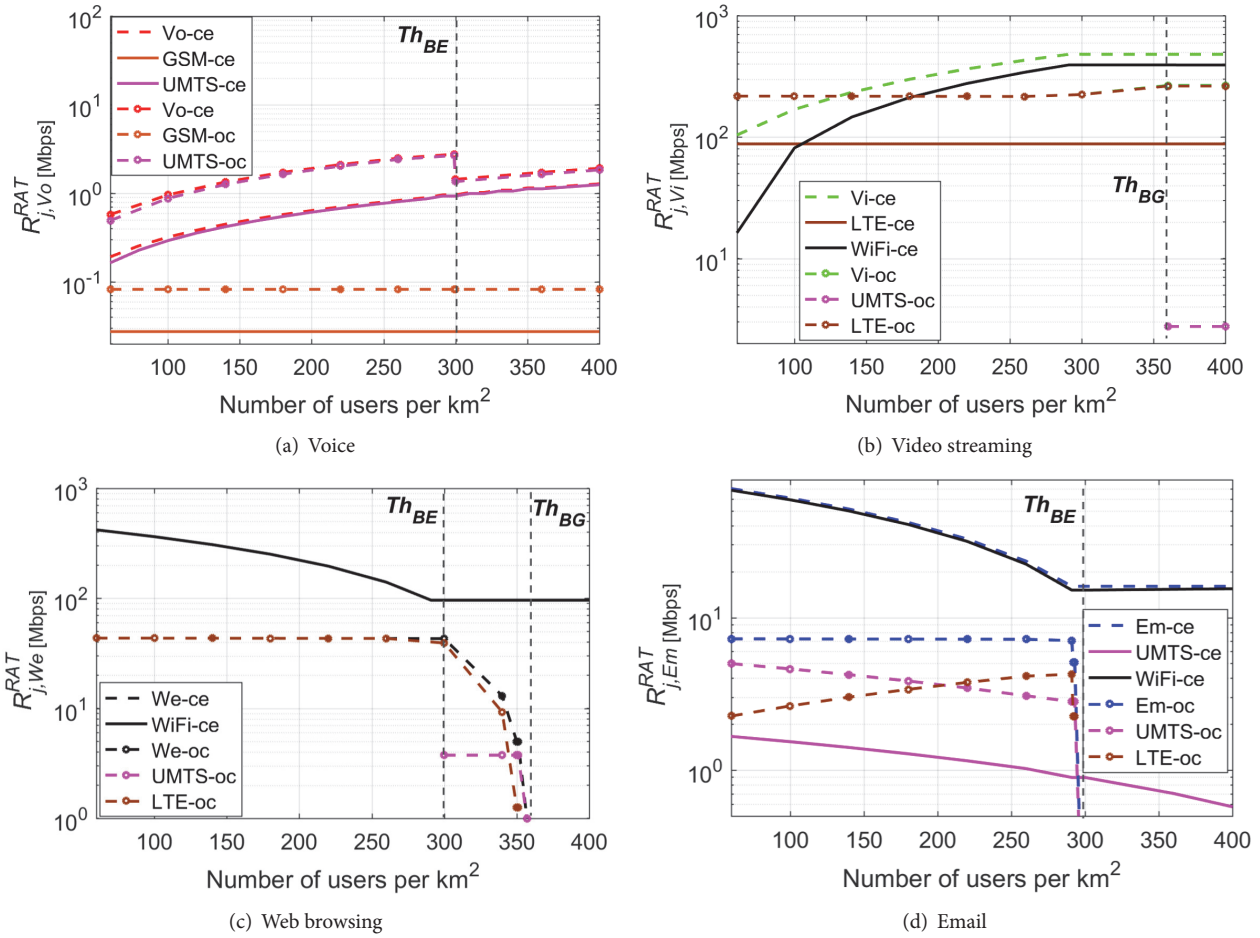
The last service is *Em*, which is shown in Figure 9(d). One can see that the demand of *ce* users is mostly responded by Wi-Fi, except for a tiny portion which is covered by UMTS. However, for *oc* users the whole requested data rate is allocated from LTE and UMTS, which are the first and second preferred available RATs, respectively. The absence of LTE as the second priority RAT in *ce* areas is due to the fact that *Em* is categorised as a BE service; therefore, it has the lowest priority, and the unallocated capacity of LTE is not sufficient to cover both demands from *ce* and *oc* users. Alternatively, a small part of *ce* users' demand is directed to UMTS as the third preferred choice for RAT selection, besides Wi-Fi.

A summary of matching RATs and services is presented in Table 6 as a comparison with the assumptions of Table 5. One can see from Table 6 that at most two RATs are assigned to a specific service, for both *ce* and *oc* areas, which are the highest priority ones, except for a small share of *Em-ce* that uses UMTS as the third priority. It is also notable that for *We-oc* and *Em-oc* users, since the first priority RAT is Wi-Fi, which is not accessible for *oc* users, they are alternatively connected to LTE and UMTS, as the second and third most suitable RATs, respectively.

6. Conclusions

This paper proposes a centralised cooperative mechanism of RRM for Virtual RANs, based on aggregation and virtualisation of all the available radio resources from different RATs. In this context, the roles of VNOs and InPs are separated, while they are interacting to maximise the overall performance efficiency. VNOs do not own the underlying infrastructure and accordingly are not dealing with the management of radio resources. Alternatively, they demand a centralised virtualisation platform, called VRRM, for *Capacity-as-a-Service* in order to satisfy their users in respect to the contracted SLAs. VRRM is in charge of managing the aggregated capacity, which is provided by InPs through the CRRM entity. On the other hand, CRRM is responsible for distributing the heterogeneous traffic among the available RATs, based on some decision criteria for defining the most suitable service-to-RAT assignment. The model is further extended to handle the extreme situations resulting from high traffic loads, when the capacity is not sufficient to provide all users with at least the minimum contracted data rate. In this case, the algorithm introduces delay to the lowest priority users, releasing enough capacity so that the rest of higher priority ones continue their service.

The main goal from the VRRM perspective is to maximise the utilisation of total capacity, which is provided by aggregating the RRUs from different RATs in order to satisfy the SLAs to the highest achievable level, while maintaining a level



(a) Voice

(b) Video streaming

(c) Web browsing

(d) Email

Figure 9: Allocation of capacity from the available RATs to each service.

of fairness among users considering the priority of services and to realise an acceptable level of isolation. To achieve this goal, VRRM has to closely interact with CRRM in order to calculate the demanded data rates of each service based on the information about the available capacity provided by CRRM, and then returning the demanded capacity to CRRM, to be provided by the available RATs. The key objective of CRRM is to associate these demands to the most suitable RATs, in order to meet the QoS requirements of each service.

The performance of the proposed model is analysed by realising a practical scenario. While all users have access to the cellular RATs, only 25% of them can use Wi-Fi. The results from the VRRM evaluation confirm that when there is no shortage of resources, the model can capture the demanded capacity of *Vo, Vi, We, and Em* users, according to the concept of proportional fairness, which is consistent with the predefined serving weights of 50, 30, 6, and 1, respectively. Moreover, all SLAs are satisfied independent of the variation of traffic load, which yields a desired level of isolation. Under extreme traffic loads in *oc* areas, the algorithm delays the users based on their service priorities, trying to serve the maximum number of high priority ones. In any case, 100% of the aggregated capacity is used. From a CRRM viewpoint, all the demanded capacities from different services are addressed by maximum two available RATs, which are the highest priority ones according to the predefined table of service-RAT matching, the only exception being a small portion of BE users' demand, since they have the lowest priority.

Data Availability

The software used for solving the optimisation problems, CVX, is publically available online at this address: http://cvxr.com/cvx/. The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is developed within the framework of the COST Action CA15104, the Inclusive Radio Communication Networks for 5G and beyond (IRACON).

References

[1] S. Glisic, *Advanced Wireless Networks: Technology and Business Models*, John Wiley and Sons, Chichester, UK, 2016.

[2] H. Wen, P. K. Tiwary, and T. Le-Ngoc, *Wireless Virtualization*, Springer, Montreal, Canada, 2013.

[3] P. Marsch, O. Bulakci, O. Queseth, and M. Boldi, *5G System Design: Architectural and Functional Considerations and Long Term Research*, John Wiley and Sons, Hoboken, NJ, USA, 2018.

[4] W. Xiang, K. Zheng, and X. Shen, *5G Mobile Communications*, Springer International Publishing, Zürich, Switzerland, 2017.

[5] C. Liang and F. Yu, "Enabling 5G mobile wireless technologies," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 218, 2015.

[6] J. Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017.

[7] A. Aijaz, "Towards 5G-enabled Tactile Internet: Radio resource allocation for haptic communications," in *Proceedings of the 2016 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, Doha, Qatar, April 2016.

[8] G. Tseliou, F. Adelantado, and C. Verikoukis, "Scalable RAN virtualization in multitenant LTE-A heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6651–6664, 2016.

[9] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, "Scheduling wireless virtual networks functions," *IEEE Transactions on Network and Service Management*, vol. 13, no. 2, pp. 240–252, 2016.

[10] M. Farooq, H. Ghazzai, E. Yaacoub, A. Kadri, and M. Alouini, "Green virtualization for multiple collaborative cellular operators," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 3, pp. 420–434, 2017.

[11] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, and Y.-D. Kim, "Fairness-aware radio resource management in downlink OFDMA cellular relay networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 5, pp. 1628–1639, 2010.

[12] X. Yu and H. Zhu, "Optimal resource management with delay differentiated traffic and proportional rate constraint in heterogeneous networks," *Journal of Communications*, vol. 9, no. 9, pp. 714–722, 2014.

[13] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[14] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 27–35, 2013.

[15] S. Singh, S. Yeh, N. Himayat, and S. Talwar, "Optimal traffic aggregation in multi-RAT heterogeneous wireless networks," in *Proceedings of the IEEE International Conference on Communications Workshops (ICC '16)*, pp. 626–631, Kuala Lumpur, Malaysia, May 2016.

[16] M. Gerasimenko, D. Moltchanov, R. Florea et al., "Cooperative radio resource management in heterogeneous cloud radio access networks," *IEEE Access*, vol. 3, pp. 397–406, 2015.

[17] S. Khatibi and L. M. Correia, "Modelling virtual radio resource management in full heterogeneous networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 73, 2017.

[18] L. Caeiro, B. Rouzbehani, and L. M. Correia, "A fair mechanism of virtual radio resource management in multi-RAT wireless het-nets," in *Proceedings of the 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–5, Montreal, QC, Canada, October 2017.

[19] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: a survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 462–476, 2016.

[20] B. Rouzbehani, L. M. Correia, and L. Caeiro, "Radio resource and service orchestration for virtualised multi-tenant mobile het-nets," in *Proceedings of the 2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–5, Barcelona, Spain, April 2018.

[21] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G radio access network slicing: radio interface protocol features and configuration," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 2–10, 2018.

[22] C. Liang and F. Yu, "Wireless network virtualization: a survey, some research issues and challenges," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 1, pp. 358–380, 2015.

[23] W. Song and W. Zhuang, *Interworking of Wireless LANs and Cellular Networks*, Springer, New York, NY, USA, 2012.

[24] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, 2013.

[25] S. Khatibi and L. M. Correia, "A model for virtual radio resource management in virtual RANs," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, pp. 1–12, 2015.

[26] G. Miao, G. Song, and G. Li, *Energy and Spectrum Efficient Wireless Network Design*, Cambridge University Press, Cambridge, UK, 2015.

[27] "CVX – Software for Disciplined Convex Programming," http://cvxr.com, Feb. 2017.

[28] S. Mehrotra, "On the implementation of a primal-dual interior point method," *SIAM Journal on Optimization*, vol. 2, no. 4, pp. 575–601, 1992.

[29] D. Marabissi and R. Fantacci, "Heterogeneous public safety network architecture based on RAN slicing," *IEEE Access*, vol. 5, pp. 24668–24677, 2017.

[30] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network Slicing Management and Prioritization in 5G Mobile Systems," in *Proceedings of the European Wireless 2016 – 22th European Wireless Conference*, Oulu, Finland, 2016.

[31] E. Hossain, *Heterogeneous Wireless Access Networks: Architectures and Protocols*, Springer, Boston, MA, USA, 2009.

[32] G. Piao, *Radio Resource Management for Integrated Services in Multi-radio Access Networks*, Springer, Kassel, Germany, 2007.

[33] H. Chen, L. Huang, S. Kumar, and C. J. Kuo, *Radio Resource Management for Multimedia QoS Support in Wireless Networks*, Springer, Boston, MA, USA, 2004.

[34] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, 2012.

[35] S. Khatibi and L. M. Correia, "Modelling of virtual radio resource management for cellular heterogeneous access networks," in *Proceedings of the 2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1152–1156, Washington DC, USA, September 2014.

[36] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, NY, USA, 4th edition, 2002.

[37] N. Mattew, O. Sadiku, and A. Warsame, *Signals and Systems: A Primer with MATLAB*, CRC Press, Boca Raton, FL, USA, 2015.

[38] P. A. Carreira, *Data Rate Performance Gains in UMTS Evolution to LTE at the Cellular Level [Msc Thesis]*, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, 2011.

[39] Mobile Cloud Networking (MCN), "Future Communication Architecture for Mobile Cloud Services," Tech. Rep., 2014, Report No. D4.3, Version 1, https://cordis.europa.eu/project/rcn/105938_en.html.

[40] T. Shuminoski and T. Janevski, "Radio network aggregation for 5G mobile terminals in heterogeneous wireless and mobile networks," *Wireless Personal Communications*, vol. 78, no. 2, pp. 1211–1229, 2014.

[41] N. Ferdosian, M. Othman, B. M. Ali, and K. Y. Lun, "Greedy–knapsack algorithm for optimal downlink resource allocation in LTE networks," *Wireless Networks*, vol. 22, no. 5, pp. 1427–1440, 2016.

[42] Ericsson., *Ericsson Mobility Report*, Stockholm, Sweden, 2017, https://www.ericsson.com/assets/local/mobility-report/documents/2017/ericsson-mobility-report-november-2017central-and-eastern-europe.pdf.