Towards Cross-Layer Optimization of Virtualized Radio Access Networks

Behnam Rouzbehani[1], Vladimir Marbukh[2], Kamran Sayrafian[2], Luis M. Correia[1]

[1]IST-University of Lisbon/INESC-ID
Lisbon, Portugal
{behnam.rouzbehani, luis.m.correia}@tecnico.ulisboa.pt

[2]National Institute of Standards & Technology
Gaithersburg, MD, USA
{vladimir.marbukh, kamran.sayrafian}@nist.gov

Abstract—**This paper proposes an approach to cross-layer optimization of virtualized Radio Access Network resources in future mobile communications. Assuming that the Virtual Network Operators (VNOs) guarantee contracted Service Level Agreements (SLAs) with the users, the proposed approach uses weighted proportional fairness as a basis for allocation of the remaining capacity. This allocation is achieved by a distributed, pricing-based solution to a two-layer convex optimization problem. Through this mechanism, some of the key functionalities of the centralized virtualization platform are transferred to the individual VNOs and users. This allows for a drastic reduction in the complexity of the system management compared to the previously proposed centralized approaches. Therefore, it leads to a much more scalable design for dense network deployments with real-time applications. Another advantage of the proposed distributed cross-layer optimization is the enhanced level of isolation among different VNOs. The proposed approach is evaluated by simulating a scenario with 3 types of VNOs and differentiated SLAs sharing radio resources from an underlying physical heterogeneous network. Results for the 4 types of service classes confirm that given sufficient aggregate capacity, all SLAs are satisfied, the entire aggregated capacity is utilized, and the residual available capacity is shared among the users proportionally fair.**

Keywords- Virtualization, multi-tenancy, service level agreements, distributed resource management

I. INTRODUCTION

Service-oriented architecture brings significant potential benefits to both users and network operators in future mobile communication. The benefits include more flexibility in resource sharing (resulting in optimized utilization of network resources), wider range of customized services to suit users' requirements, as well as reduction in the CAPEX/OPEX costs [1]. The one-size-fits-all architecture is not likely to be appropriate solution for a diverse range of services. Therefore, wireless network *virtualization* has been proposed as an alternative technology to support service-oriented architectures [2]. This involves decoupling of the services and functionalities from the underlying Radio Access Networks (RANs). Virtualization transforms the physical infrastructure into multiple logical network instances that are shared among different *tenants,* i.e., Virtual Network Operators (VNOs). VNOs are expected to operate in an isolated manner. This implies some changes in the traditional role of Mobile Network Operators (MNOs). In contrast to MNOs , VNOs do not own the infrastructure. Instead, they obtain the capacity from a centralized virtualization platform and enforce their own service requirements and policies in the process of Radio Resource Management (RRM) [3].

RRM is one of the most important functionalities in mobile networks that directly impacts the users' Quality of Service (QoS). Given the diverse range of possible services which might require individual management, RRM could be extremely challenging to implement [4]. Resource slicing, enabled by virtualization technology, is a solution that can address the specific requirements of each service by providing a certain degree of performance isolation. This will make sure that regardless of the variation of different parameters in the network (e.g., traffic load or channel condition), the desired efficiency level of independent slices can always be achieved [5]. RRM in virtualized Heterogeneous Networks (Het-Nets) should not only optimize the performance of various slices but also maximize the utilization of the overall shared resources [6]. To the best of the authors' knowledge, there are no comprehensive studies that thoroughly cover key issues such as differentiated service provisioning, performance isolation, and fairness for RRM in virtualized Het-Nets.

Centralized approaches to RRM, considered in the literature, suffer from scalability limitation, especially when it comes to low-latency real-time applications' requirements [7], [8]. Therefore, there is a need for decentralized RRM approaches in future mobile communication networks. In [9], a distributed RRM model based on non-cooperative game theory is proposed for dense wireless 5G networks where each Base Station (BS) tries to maximize its payoff. While this model achieves energy efficiency, it does not consider the customized specifications and requirements of different services. An adaptive two-layer decentralized RRM with slow and fast timescales for adaptation of the central manager and users has been proposed in [10] for 5G networks. However, [10] has not considered network virtualization and slicing concepts, which are key enablers of 5G. Another distributed RRM approach with a focus on multi-connectivity in 5G has been described in [11]. While the proposed approach aims at reducing the processing costs and signalling overhead, it lacks the notion of RAN slicing, isolation, as well as service orientation in the model development.

This paper introduces a distributed approach for RRM to overcome the scalability problem of the centralized solution discussed in [12, 13]. Decentralization is achieved through the use of a two-stage distributed optimization with pricing

adaptation on a *fast* and *slow* time scales [14, 15]. At the faster time-scale, and assuming that VNO capacities do not change, users adjust their rates based on the congestion pricing. At the slower time scale, each VNO adjusts its own capacity according to its assigned congestion price, subject to the total aggregate capacity of the system. This decentralization takes advantage of the dual role of congestion prices used for both adjustment of the rates by elastic users and capacity expansion/reduction.

The rest of the paper is organized as follows. Section II describes the system architecture and the main assumptions. Section III proposes the optimization framework for our decentralized approach. Section IV describes a case study scenario for performance evaluation through simulation. Section V reports and analyzes results for the case study. Finally, concluding remarks and future plans are summarized in section VI.

II. System Architecture and Assumptions

Figure 1 displays the mechanism of service-oriented RAN slicing and resource management, and interaction of different entities in the system. The Virtual-RRM (VRRM) entity is responsible for configuring the RAN protocol stack and QoS metrics according to the slice requirements. Those requirements are enforced by different VNOs considering their specific policies. As an example, assume that VNOs *A* and *B* provide two types of services with different requirements. For the *slice A* with high throughput requirements, radio flow *A*, which corresponds to a customized radio bearer is configured to support multi-connectivity. Therefore, slice *A* is using the resources from 2 different radio access points. On the other hand, the network *slice B* is configured with only one connection according to the provided policy.

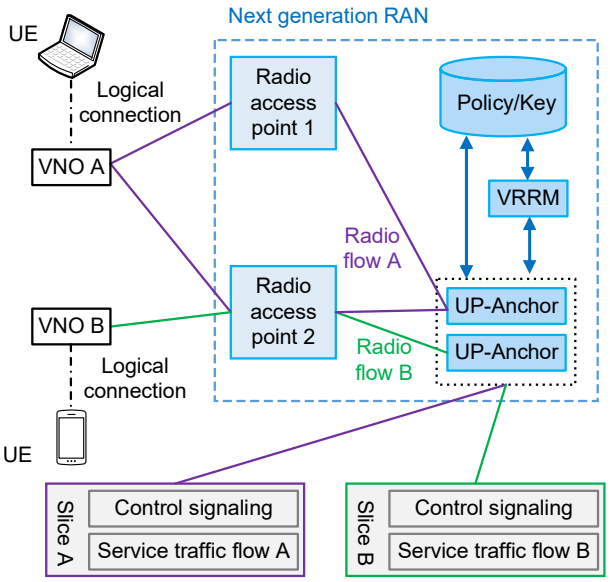

Figure 1. Service-oriented RAN slicing

The User Plane Anchor (UP-Anchor) distributes the traffic flow in each slice. Accordingly, a RAN slice is composed of a separated control plane to address the policies of VNOs, enforced by VRRM, and a data plane to carry the service traffic.

The required capacity delivery is subject to the SLA agreements between the VNOs and users. We consider the following three categories of SLA contracts [12]:

- Guaranteed Bitrate (GB): This is the highest priority category for which minimum and maximum thresholds for data rate assignment must be always guaranteed regardless of the variation of traffic load and network status.
- Best effort with minimum Guaranteed (BG): This is the second highest priority category, for which a minimum level of data rate is guaranteed. Higher data rates are served in a best effort manner if available.
- Best Effort (BE): This is the lowest priority category, for which there is no level of service guarantees and users are served in a pure best effort manner.

III. Optimization Framework

Decentralized optimization is achieved through congestion pricing at different timescales. Subsection A describes optimization at the "fast" timescale, when *users* adjust their rates based on the congestion pricing policies (assuming fixed VNOs capacities). Subsection B describes optimization at the "slow" timescale, when each *VNO* adjusts its capacity according to its assigned congestion price subject to the total network aggregate capacity.

A. Users Rate Adaptation

Assume that VRRM has already distributed the total aggregated capacity of R_{VRRM} among N_V number of VNOs, such that the capacity share of each one is equivalent to $C_v^{VNO}, v \in \{1,2,\dots,N_V\}$. Accordingly, each VNO allocates the rate of R_i^{Usr} to each of its connected users such that:

$$\sum_{i\in I_v} R^{Usr}_{i[\text{Mbps}]} \le C^{VNO}_{v[\text{Mbps}]}, \qquad (1)$$

where there are N_s groups of users per VNO v, representing N_s different service slices served by that VNO, forming a set of total users, I_v, to be served by the VNO v. It is further assumed that the N_s service slices do not overlap, i.e., each user performs one service and belongs to a specific VNO:

$$I_v = \left\{I_{v_1}, I_{v_2}, \dots, I_{v_{N_s}}\right\}, \qquad I_{v_1} \cap I_{v_2} = \emptyset. \qquad (2)$$

For each user $i \in I_{v_j}, j = \{1,2,\dots,N_s\}$, individual net utility functions $U_i\ (R_i^{Usr})$ are introduced in the form of logarithmic objectives to capture the required rate of R_i^{Usr} according to the criterion of weighted proportional fairness:

$$U_i\left(R^{Usr}_{i[\text{Mbps}]}\right) = \lambda_i\, log\left(\frac{R^{Usr}_{i[\text{Mbps}]}}{R^{Ref}_{[\text{Mbps}]}}\right) - p_v \frac{R^{Usr}_{i[\text{Mbps}]}}{R^{Ref}_{[\text{Mbps}]}}, \qquad (3)$$

where $\lambda_i > 0$ values are tuning parameters to prioritize the service slices and p_v is price of a unit bandwidth offered by the VNO v. R^{Ref} is a reference value to normalize the data rates. It is also assumed that a user's assigned data rate is positive and lies within a given interval of *minimum* and *maximum* thresholds:

$$0 \le R^{Usr_{min}}_{i[\text{Mbps}]} \le R^{Usr}_{i[\text{Mbps}]} \le R^{Usr_{max}}_{i[\text{Mbps}]}, \qquad i \in I_v. \qquad (4)$$

By solving the individual convex optimization problem $\max_{R_i>0} U_i(R_i)$ subject to constraint (4), each user i calculates its rate as follows:

$$R^{Usr}_{i[\text{Mbps}]}(p_v) = \begin{cases} \lambda_i/p_v & if \quad R^{Usr_{min}}_{i[\text{Mbps}]} \le \lambda_i/p_v \le R^{Usr_{max}}_{i[\text{Mbps}]} \\ R^{Usr_{min}}_{i[\text{Mbps}]} & if \quad \lambda_i/p_v < R^{Usr_{min}}_{i[\text{Mbps}]} \\ R^{Usr_{max}}_{i[\text{Mbps}]} & if \quad \lambda_i/p_v > R^{Usr_{max}}_{i[\text{Mbps}]} \end{cases} \quad (5)$$

This solution, which is shown in Figure 2, is determined by condition that the slope of user utility coincides with the current congestion price within the domain of $\left[R_i^{Usr_{min}}, R_i^{Usr_{max}}\right]$. Note that due to the lower bound constraints in (4), the problem may not have a feasible solution. Therefore, there is a need for an admission control mechanism to ensure the existence of a feasible solution, i.e.,

$$\sum_{i\in I_v} R^{Usr_{min}}_{i[\text{Mbps}]} \le C^{VNO}_{v[\text{Mbps}]}, \qquad v = \{1,2,\dots,V\}. \quad (6)$$

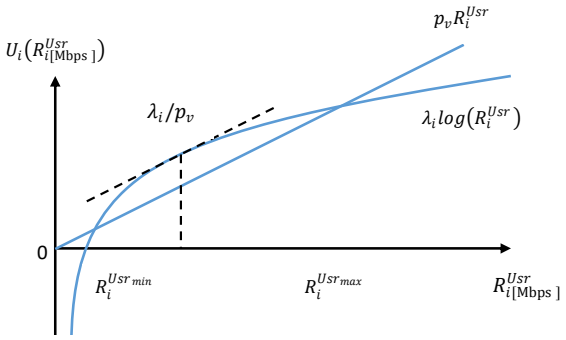

Figure 2. Individual user's rate optimization.

The optimal prices p_v^{opt}, that maximize the utilization of the VNOs' available bandwidth, are determined by the following distributed adaptive algorithm. The algorithm proceeds in discrete steps $k = \{1,2,\dots\}$. At each step k, users solve the individual optimization problems resulting in (5). If (1) is satisfied, i.e., the aggregate data rate of the users does not exceed the total capacity of the associated VNO, then in step $k+1$ the price $p_{v,k+1}$ is reduced to motivate users to buy more bandwidth. However, if the constraint (1) is not satisfied, $p_{v,k+1}$ is increased, resulting in a decrease of users' data rates. The main idea here is to maximize utilization of the available bandwidth in an efficient way. The price adaptation model can be expressed as follows [15]:

$$p_{v,k+1} = \left[p_{v,k} + h\frac{1}{R^{Ref}_{[\text{Mbps}]}}\left(\sum_{i\in I_v} R^{Usr}_{i[\text{Mbps}]} - C^{VNO}_{v[\text{Mbps}]}\right)\right]^+, \quad (7)$$

where $[x]^+ = \max(0,x)$ and $h > 0$ is a small positive constant which regulates the tradeoff between optimality under stationary scenario and adaptability under non-stationary scenario, e.g., due to changing set of users.

The main advantage of this approach is that VNOs do not have to know users' utilities, which are considered private information.

B. VNOs' Capacity Optimization

In the slower time-scale each VNO adjusts its own capacity by negotiating the price with VRRM. Assuming that the total available capacity of VRRM is R^{VRRM}, the adaptation of capacities among the tenant VNOs, C_v^{VNO} is subject to the total VRRM capacity:

$$\sum_{v=1}^{N_v} C^{VNO}_{v[\text{Mbps}]} \le R^{VRRM}_{[\text{Mbps}]}. \quad (8)$$

The average price of a unit of bandwidth in the entire system at step $m = \{1,2,\dots\}$ is as follows:

$$P_m^{ave} = \frac{1}{R^{VRRM}_{[\text{Mbps}]}}\sum_{v=1}^{N_v} C^{VNO}_{v[\text{Mbps}]}\, P_{v,m} \quad (9)$$

where $P_{v,m}$ is the price of a unit of bandwidth assigned to VNO v from VRRM at step m.

We propose the following capacity adaptation algorithm for the VNOs according to [15]:

$$C^{VNO}_{v,m+1[\text{Mbps}]} = \begin{cases} C^{VNO_{min}}_{v[\text{Mbps}]} & if \quad C^{VNO}_{v,m[\text{Mbps}]} \le C^{VNO_{min}}_{v[\text{Mbps}]} \\ C^{VNO_{max}}_{v[\text{Mbps}]} & if \quad C^{VNO}_{v,m[\text{Mbps}]} > C^{VNO_{max}}_{v[\text{Mbps}]} \\ C^{VNO}_{v,m[\text{Mbps}]} + H\left(P_{v,m} - P_m^{ave}\right) & otherwise \end{cases} \quad (10)$$

where $H > 0$ is a small constant. The minimum and maximum capacities of VNO are respectively:

$$C^{VNO_{min}}_{v[\text{Mbps}]} = \sum_{i\in I_v} R^{Usr_{min}}_{i[\text{Mbps}]}, \quad v = \{1,2,\dots V\} \quad (11)$$

$$C^{VNO_{max}}_{v[\text{Mbps}]} = \sum_{i\in I_v} R^{Usr_{max}}_{i[\text{Mbps}]}, \quad v = \{1,2,\dots V\} \quad (12)$$

Algorithm (10)-(12) increases (decreases) the capacity of a VNO if its corresponding price is higher (lower) than the average price (9).

IV. Case Study

Consider an area that is uniformly covered by GSM, UMTS, LTE and Wi-Fi RATs according to the coverage plan and the specifications in [12]. To estimate the available aggregate capacity of VRRM, we employ the convolution Probability Density Function (PDF) function specified in [12], and assume that there is upper and lower bounds on the data rate of each Radio Resource Unit (RRU). We also assume that users experience independent channel fading with a Rayleigh Distribution. By randomly selecting from the proposed convolution of all the RRUs' PDFs, the VRRM capacity is obtained to be 590 Mbps.

Network parameters are specified in Table 1. As observed, it is assumed that 3 VNOs with different SLA types, i.e., GB, BG and BE, provide 4 different service classes: *Conversational* (Con), *Streaming* (Str), *Interactive* (Int.) and *Background* (Bac.). VNO GB delivers Voice (Voi), Video calling (Vic), Video streaming (Vis) and Music streaming (Mus). VNO BG

serves File sharing (Fil), Web browsing (Web) and Social Networking (Soc) services, while VNO BE provides Smart metering (Sma) and Email (Ema).

Table 1 – Network parameters

VNO	Service	Class	$R_{i[Mbps]}^{Usr}$	$U_{[\%]}^{srv}$	λ_i	SLA
1	Voi	Con.	[0.032, 0.064]	25	5	GB
	Vic		[1, 4]	15	4	
	Vis	Str.	[2, 13]	45	3	
	Mus		[0.064, 0.32]	15	1	
2	Fil	Int.	$[1, R^{VRRM}]$	50	4	BG
	Web		$[0.5, R^{VRRM}]$	35	3	
	Soc		$[0.4, R^{VRRM}]$	15	2	
3	Sma	Bac.	$[0, R^{VRRM}]$	25	4	BE
	Ema		$[0, R^{VRRM}]$	75	4	

R_i^{Usr} represents the range of customized service data rates according to the internal policy of each VNO and $U_{[\%]}^{srv}$ is the traffic mix of each service. Moving from top to bottom in Table 1, the priority of VNOs and their corresponding services decrease. This is due to the SLA types and decrease in the values of serving weights λ_i.

V. Analysis of the Results

The impact of traffic load variation on the capacity share of VRRM is shown in Figure 3. The *dotted* lines represent the minimum demands of VNOs GB and BG, and the *dashed* line represents the maximum demand of VNO GB. As the number of users increases, the capacity share of VNO GB increases due to its highest priority. At the same time, the capacity of the other two VNOs decreases while satisfying the minimum demanded data rates (assuming there is no capacity shortage).

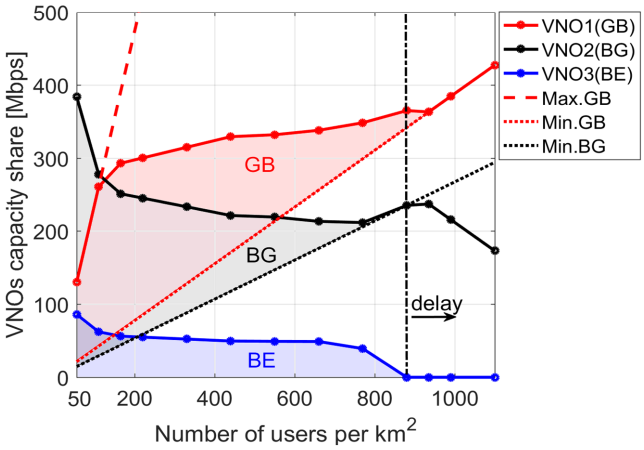

Figure 3 - Capacity share of VRRM among the VNOs.

When there is not enough capacity to serve all users with the minimum demanded data rates, VRRM starts reducing the capacity share of lower priority VNOs to compensate for the capacity need of the higher priority VNOs. Accordingly, those low priority VNOs also start delaying the users with the lowest service priority. For example, as observed in Figure 3, before

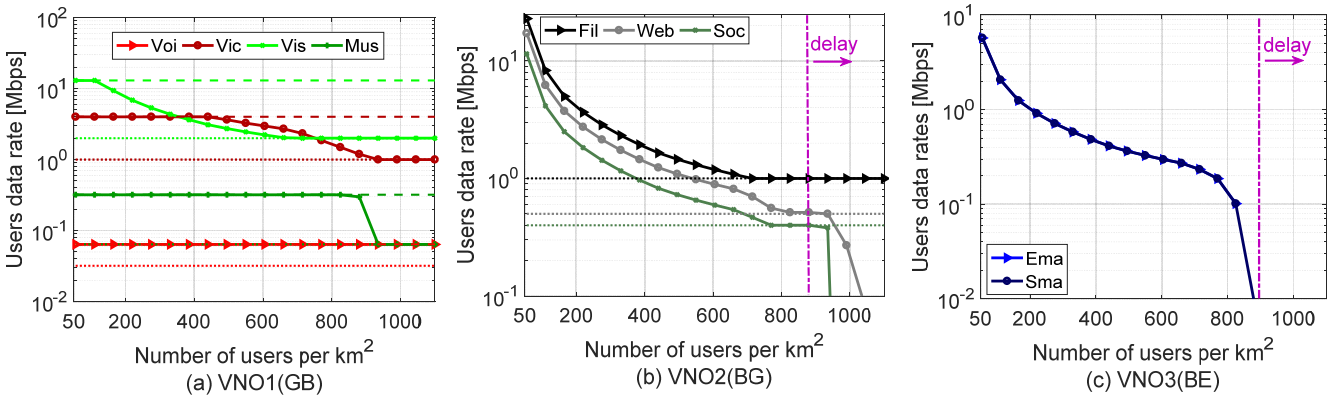

Figure 4. Average data rate of served users, in the three VNOs.

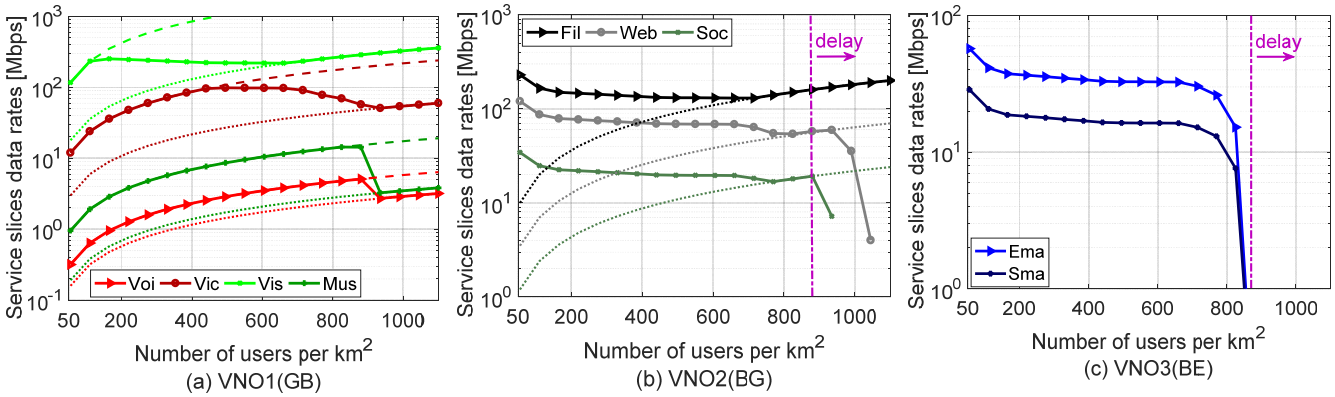

Figure 5. Share of the available bandwidth by service and VNO.

any delay starts for VNO BG users (i.e. the point when the capacity share of VNO BG drops to a value less than the minimum demand), the capacity of VNO BE has decreased to zero. This indicates that all BE users have already been delayed.

After capacity allocation to all VNOs, the capacity shares of each VNO among its connected users and service slices in the second management layer are presented in Figures 4 and Figure 5 respectively. Looking at the VNO GB, the users' data rates always vary between the minimum and maximum thresholds predefined in Table 1. As delays start for the users of VNO BG (i.e., when the number of users is roughly around 900), the data rates of all the GB users have already reached down to their minimum acceptable threshold.

These results confirm that when there is no constraint, the share of data rate among the users is exactly *proportional* to their serving weights. For example, before the users' data rates drop to the minimum, the capacity of VNO BG is shared among Fil, Web and Soc users proportional to their service weights: 4, 3 and 2, respectively. It is also observed that when there is not enough capacity to serve all users, VNO BG first delays Soc users followed by Web users in order to provide enough capacity for all the Fil users.

The results for VNO BE (shown in Fig. 4.c) confirm that the users' data rates for both Ema and Sma services are similar as they have the same priority, and range of specified data rates. However, the total allocated capacity to the service slice of Ema is higher than that of Sma. This is because the number of Ema's users is three times greater than Sma's users.

VI. Conclusions and Future Work

This paper proposes a pricing-based distributed convex optimization model for radio resource management in virtual wireless Het-Nets comprised of different access technologies. A two-layer optimization problem with *slow* and *fast* price adaptation mechanism is developed for the VRRM and VNOs as well as for the VNOs and end-users in order to achieve a higher level of isolation and privacy. Further, this methodology significantly contributes to the reduction of complexity compared to the centralized approaches. Therefore, it will facilitate dense deployment of real-time applications.

In order to evaluate the model, a scenario with 3 different SLA types has been considered. The VNOs share the total aggregated capacity of the underlying physical RATs to satisfy the QoS requirements of different service slices for four classes of services. Simulation results confirm that in a case when system has sufficient capacity to satisfy all SLAs, the proposed algorithm (a) ensures that all SLAs are satisfied and (b) allocates the entire remaining system capacity proportionally fair with predetermined weights. In case of capacity shortage, the admission control process delays the lowest priority users in order to provide necessary capacity for the rest of the users to continue their service at the minimum acceptable rates. Following this methodology, the cooperation between different entities will result in 100% usage of the system capacity.

The authors plan to investigate the possibility of distributed admission control and its impact on the proposed model. In addition, research is underway to evaluate the effect of some of the simplifying assumptions such as separation of time scales for users data rates and VNOs capacities adaptations, as well as delays in the pricing information.

Acknowledgment

Authors would like to acknowledge the COST CA15104 (IRACON) for the productive technical discussions that resulted into this collaborative work.

References

[1] Z. Chang, Z. Zhou, S. Zhou, T. Chen and T. Ristaniemi, "Towards Service-Oriented 5G: Virtualizing the Networks for Everything-as-a-Service", *IEEE Access,* Vol. 6, Dec. 2017, pp. 1480–1489.

[2] Z. Feng, L. Ji, Q. Zhang and W. Li, "A Supply-Demand Approach for Traffic-Oriented Wireless Resource Virtualization with Testbed Analysis", *IEEE Transactions on Wireless Communications,* Vol. 16, No. 9, Jun. 2017, pp. 6077–6090.

[3] M. Elkhodr, Q.F. Hassan and S. Shahrestani, *Networks of the Future: Architectures, Technologies, and Implementations*, CRC Press, Boca Raton, FL, USA, 2018.

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

[5] J. Lucena, P. Ameigeiras and D. Lopez, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges", *IEEE Communications Magazine*, Vol. 55, No. 5, May 2017, pp. 80–87.

[6] A. Aijaz, "Towards 5G-enabled Tactile Internet: Radio Resource Allocation for Haptic Communications", in *Proc. of WCNC'16 - 17th IEEE Wireless Communications and Networking Conference*, Doha, Qatar, Apr. 2016.

[7] S. Singh, S. Yeh, N. Himayat, S. Talwar, "Optimal Traffic Aggregation in Multi-RAT Heterogeneous Wireless Networks", in *Proc. of ICC'16 – 52th IEEE International Conference on Communications*, Kuala Lumpur, Malaysia, May 2016.

[8] M. Gerasimenko, D. Moltchanov and R. Florea, "Cooperative Radio Resource Management in Heterogeneous Cloud Radio Access Networks", *IEEE Access*, Vol. 3, Apr. 2015, pp. 397–406.

[9] P. Sroka and A. Kliks, "Playing Radio Resource Management Games in Dense Wireless 5G Networks", *Hindawi Journal of Mobile Information Systems*, Vol. 2016, Nov. 2016, pp. 1 – 18.

[10] F. Teng and D. Guo, "Resource Management in 5G: A Tale of Two Timescales", in *Proc. of ACSSC'15 - 49th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA.

[11] V. Monteiro, D. Sousa and T. Maciel, "Distributed RRM for 5G Multi-RAT Multi-connectivity Networks", *IEEE Systems Journal* (Early Access), Jun. 2018, pp. 1 – 13.

[12] B. Rouzbehani, L.M. Correia and L. Caeiro, "Radio Resource and Service Orchestration for Virtualised Multi-Tenant Mobile Het-Nets", in *Proc. of WCNC'18 – 19th IEEE Wireless Communications and Networking Conference,* Barcelona, Spain, Apr. 2018.

[13] B. Rouzbehani, L.M. Correia and L. Caeiro, "A Fair Mechanism of Virtual Radio Resource Management in Multi-RAT Wireless Het-Nets", in *Proc. of PIMRC'17 – 28th IEEE Symposium on Personal, Indoor and Mobile Radio Communications,* Montreal, QC, Canada, Oct. 2017.

[14] J. K. MacKie-Mason and H.R. Varian, "Pricing congestible network resources," *IEEE Journal of Selected Areas in Communications*, Vol. 13, No. 7, Sep. 1995, pp. 1141 – 1149.

[15] X. Lin, N.B. Shroff, and R. Srikant, "A Tutorial on Cross-Layer Optimization in Wireless Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 24, No. 8, Aug. 2006, pp 1452 – 1463.