Utility Proportional Resource Allocation for Users with Diverse SLAs in Virtualized Radio Access Networks

Behnam Rouzbehani[1], Vladimir Marbukh[2], Kamran Sayrafian[2]

[1]IST-University of Lisbon/INESC-ID
Av. Rovisco Pais 1, 1049-001
Lisbon, Portugal

[2]National Institute of Standards & Technology
100 Bureau Drive, Stop 8910
Gaithersburg, MD, USA

Abstract **– A virtualization platform is responsible for allocation and aggregation of radio resources from different access technologies as well as the distribution of the total capacity among Virtual Network Operators (VNOs). The Radio Resource Management (RRM) employed by each VNO should comply with the requirements specified in the Service Level Agreements (SLAs) of each user. A joint admission control and resource management scheme based on proportionally fair rate allocation among different users was proposed in our previous publication. Although, all SLAs are satisfied in that scheme, users with vastly different QoS requirements might not necessary be treated fairly in terms of the allocated rates. This is especially the case when the available capacities of the VNOs cannot support the maximum requested rates for all such users. This paper attempts to overcome this weakness by replacing the proportional fairness strategy with a more general concept of utility-proportional fairness. The proposed approach is evaluated by simulations under increasing congestion scenarios and the results show improved fairness in the allocated rates.**

Keywords – Virtualization, distributed resource allocation, utility-proportional fairness, Service Level Agreements (SLA), admission control.

I. Introduction

S*ervice-oriented* architecture is expected to enable flexibility in sharing and utilization of network resources, wider range of customized services, along with a reduction in the capital and operational expenditures [1]. *Virtualization* supports service-oriented architecture through decoupling of the services and functionalities from the underlying Radio Access Networks (RANs) [2]. It enables the transformation of the physical infrastructure into multiple logical networks that can be shared among different Virtual Network Operators (VNOs). As such, VNOs do not need to own the infrastructure. Instead, they obtain the resources from a centralized virtualization platform and enforce their own service requirements and policies through the process of Radio Resource Management (RRM) [3].

The diversity in users' Quality of Service (QoS) requirements drives the emergence of resource slicing along with virtualization [4]. Performance optimization in virtualized Heterogeneous Networks (Het-Nets) not only optimizes the performance of various slices but also maximizes the utilization of the overall shared resources [5]. Scalability limitations of centralized RRM necessitate decentralized resource management [6], [7]. Authors in [8] have proposed a distributed RRM model for dense 5G networks based on non-cooperative game theory. While their approach achieves energy efficiency, it does not incorporate customized specifications and requirements of different services. An adaptive two-layer decentralized RRM with slow and fast timescales has also been presented in [9]; however, the methodology does not include network virtualization and slicing concepts. Authors in [10] propose another distributed RRM with a focus on multi-connectivity in 5G networks. Their approach aims at reducing the processing costs and signalling overhead; but does not consider the notion of RAN slicing, isolation, as well as service orientation.

In our previous publication, we proposed a joint admission control and RRM for virtualized RANs [1]. Here, we extend the proportionally fair rate allocation scheme in [11] to a more general utility-proportional rate allocation [12]. The intention is to address some of the observed shortcomings of proportionally fair allocation such as giving advantage to users with low bandwidth requirements [12]. Similar to [1], our proposed scheme maximizes the aggregate system utility using a two-stage distributed optimization on a *fast* and *slow* time scale and overcomes the scalability issues of the centralized RRM [13], [14]. At the faster time scale, and given the capacities of each VNO, 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. The admission strategy, which requires limited degree of centralization, ensures system ability to guarantee minimum bandwidth requirements to the newly admitted user as well as to all users already present in the system.

The rest of this paper is organized as follows. Section II describes the system architecture and quantifies user preferences. Section III formulates system performance model. Section IV outlines the resource management scheme. Section V describes simulation scenarios and results. Finally, conclusion and plans for future research are discussed in section VI.

II. System Model

System architecture and quantification of user preferences by their corresponding utilities are described in the following subsections.

A. System Architecture

Figure 1 shows the mechanism of service-oriented RAN slicing and resource management along with interaction of different entities in the system. The Virtual-RRM (VRRM) module is a centralized virtualization platform which is responsible for configuring the RAN protocol stack and QoS metrics according to the slice requirements. Those requirements are enforced by different VNOs based on their specific policies. As an example, assume that VNOs A and B provide two types of services with different requirements. For *slice* A with high throughput requirements, radio flow A 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.

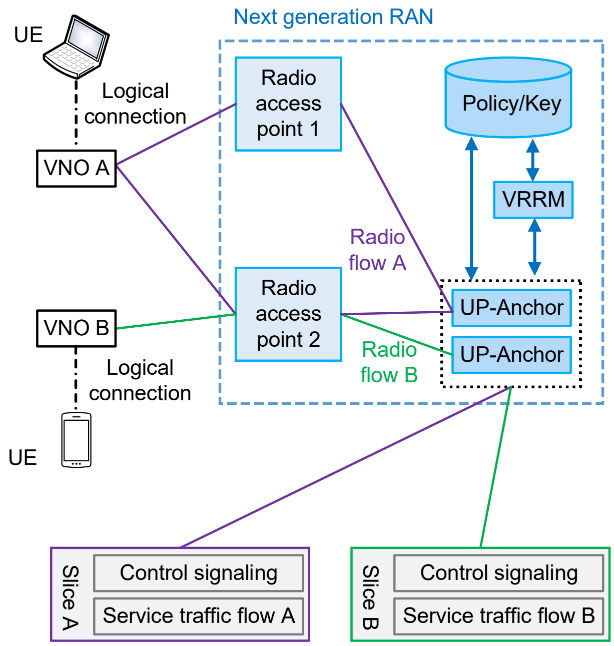

Figure 1. Service-oriented RAN slicing

The User Plane Anchor (UP-Anchor) is responsible for distributing the traffic flow in each slice. A RAN slice is composed of a control plane and a separate data plane. The required capacity allocation is subject to the SLA agreements between the VNOs and users. In this paper, we consider the following three categories of SLA contracts:

- Guaranteed Bitrate (GB): This is the highest priority category where a minimum threshold for data rate assignment must always be guaranteed regardless of the traffic load variation and network status. In addition, the assigned data rate need not exceed a maximum threshold for this SLA category.
- 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.

B. User Preferences

We assume that preference of each user for rate R can be quantified by the utility function $U(R)$, $R > 0$. In [1], we assumed the following logarithmic utility:

$$U(R) = \lambda \log(\alpha R), \tag{1}$$

However, logarithmic utility is typically inadequate for users with diverse QoS requirements which is the main driver of emerging resource slicing technology [12]. This utility function basically favors users with low bandwidth requirements, as observed in [12]. Logarithmic utility may also lead to negative utility values which could potentially cause undesirable oscillations during the rate allocation process. Some of these issues including negative utility values can be avoided by replacing utility (1) with the following utility function, shown in Fig.2.

$$U(R) = \lambda \log(1 + \alpha R). \tag{2}$$

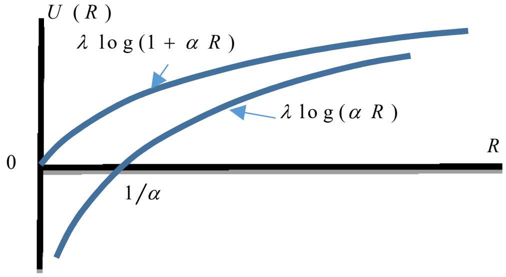

Figure 2. Logarithmic utilities

Fig. 3 exhibits the general utility of a user which requires certain minimum rate $R^{\min}$ and does not significantly benefit from rates above $R^{\max}$.

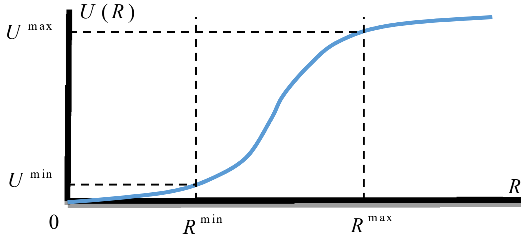

Figure 3. Utility of user with min/max rate guarantee

For example, GB users can be described by sigmoid utility $u(R)$ which is near zero for $R < R^{\min}$, quickly increases for $R^{\min} < R < R^{\max}$, and levels off for $R > R^{\max}$. A natural SLA for user with this utility is $R^{\min} < R < R^{\max}$. Sigmoid utility is often represented by the following function:

$$U(R) = \frac{k}{1 + \exp[-\alpha(R - r)]}, \tag{3}$$

where parameters $k, \alpha, r > 0$ can be expressed in terms of $R^{\min}$, $R^{\max}$, and $U^{\max} = U(R^{\max})$ [15]. Similarly, BG users are described by an utility which is near zero for $R < R^{\min}$, quickly increases for $R^{\min} < R < R^{\max}$, and then logarithmically increases for $R > R^{\max}$.

III. System Performance Optimization

The concept of utility proportional fairness and the class of utility functions which describes the SLAs considered in this paper are briefly discussed in the following subsections.

A. Utility Proportional Fairness

Let I_{sv} be the set of users obtaining service from the slice $s = 1,..,S$ of VNO $v = 1,..,V$, where sets I_{sv} with different (s,v) do not overlap, e.g., $I_{km} \cap I_{ln} = \emptyset$ if $(k,m) \neq (l,n)$, $k,l \in \{1,..,S\}$ and $m,n \in \{1,..,V\}$. Following Network Utility Maximization (NUM) framework [12], we assume that the goal of system management is maximization of the aggregate utility

$$U_\Sigma(R_i) = \sum_{s=1}^{S}\sum_{v=1}^{V}\sum_{i\in I_{sv}} U_i(R_i) \tag{4}$$

over vector of rates (R_i) allocated to users $i \in I_{sv}$, $s = 1,..,S$; $v = 1,..,V$. This maximization is a subject to the following capacity and contractual constraints. The total capacity allocated to all users serviced by VNO v, $i \in I_{sv}$, $s = 1,..,S$ cannot exceed the VNO v capacity C_v:

$$\sum_{s=1}^{S}\sum_{i\in I_{sv}} R_i \le C_v, v = 1,..,V. \tag{5}$$

Also, the aggregate capacity allocated to all VNOs cannot exceed the total system capacity C^{VNNO}:

$$\sum_{v=1}^{V} C_v \le C^{VNNO}. \tag{6}$$

The above constraints are due to data rate guarantees to a user $i \in I_{sv}$ in slice s, i.e. R_s^{min} and R_s^{max} respectively:

$$0 \le R_s^{min} \le R_{svi} \le R_s^{max}, \quad s = 1,..,S, \quad v = 1,..,V \tag{7}$$

The second set of constraints is due to guarantees on the minimum capacity of each VNO v, $C_v^{min} \ge 0$:

$$C_v \ge C_v^{min}, v = 1,..,V. \tag{8}$$

Here, we consider a distributed solution to the aggregate utility (2) maximization:

$$\max_{(C_v)}\max_{(R_i)} \sum_{s=1}^{S}\sum_{v=1}^{V}\sum_{i\in I_{sv}} U_i(R_i) \tag{9}$$

subject to constraints (5)-(8). Note that due to lower bounds in (7) and (8), optimization problem (5)-(9) may not have a feasible solution. This possibility necessitates an admission control similar to the process described in [1].

For concave user utilities, including logarithmic utilities (1) and (2), optimization problem (5)-(9) is convex; and therefore, the local maximum is also a global maximum (R_i^*). This is assuming that feasible sets (5)-(8) are non-empty. In the particular case of logarithmic utility (1), solution (R_i^*) is proportionally fair for any feasible allocation (R_i), i.e.

$$\sum_{i\in I_{sv}} \lambda_i (R_i - R_i^*)/R_i^* \le 0. \tag{10}$$

For non-concave user utilities (e.g., sigmoid utility (3)), optimization problem (5)-(9) is non-convex; and therefore, not generally tractable.

To resolve problems with proportional fairness, utility proportional fairness has been proposed in [12]. Rate allocation (R_i^*) is utility $u_i(R_i)$ proportional if

$$\sum_{i\in I_{sv}} [(R_i - R_i^*)/u_i(R_i^*)] \le 0 \tag{11}$$

for any feasible allocation (R_i). Proportional fairness (10) is a particular case of utility $u_i(R) = \lambda_i^{-1} R$ proportional fairness. It is known that utility $u_i(R)$ proportional fairness is equivalent to NUM with modified utility [12]

$$U_i(R) = \int_{R_i^{\min}}^{R} dr/u_i(r), \ R_i^{\min} \le R \le R_i^{\max}, \tag{12}$$

i.e., equivalent to aggregate utility maximization

$$\max_{(C_v)}\max_{(R_i)} \sum_{s=1}^{S}\sum_{v=1}^{V}\sum_{i\in I_{sv}} U_i(R_i). \tag{13}$$

subject to constraints (5)-(8).

Note that the NUM problem should be solved every time the set of users changes due to user arrivals/departures. Assuming that resource optimization occurs on a faster time scale changes in the number of users, distributed solution to NUM (5)-(8), (12)-(13) is discussed in the next section. As an example, consider the following utility function,

$$u(R) = [u^{\min} + \lambda^{-1}(R - R^{\min})]_{u^{\min}}^{u^{\max}}, \tag{14}$$

where $\lambda = (R^{\max} - R^{\min})/(u^{\max} - u^{\min})$, and $[z]_a^b = \max\{a, \min\{z,b\}\}$. Using utility (14) in equation (12) will lead to the following $U(R)$, and the comparison is shown in Figure 4.

$$U(R) = \left[\lambda \log\left(1 + \left(\frac{u^{\max}}{u^{\min}} - 1\right)\frac{R - R^{\min}}{R^{\max} - R^{\min}}\right)\right]_0^{U^{\max}}, \tag{15}$$

where $U^{\max} = \lambda \log(u^{\max}/u^{\min})$.

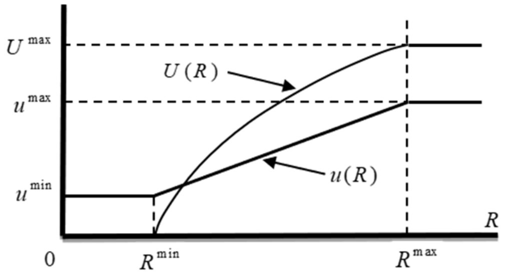

Figure 4. Piece-wise linear utility fairness

Our selection of utility (14) is due to its ability to describe BE users as well as users with rate guarantees.

IV. Resource Management

User rate and VNO capacity adaptation algorithms, given that the optimization problem (5)-(8), (12)-(13) has a feasible solution for the set of users, i.e., system has sufficient capacity to satisfy minimum rate requirements for all users in the system is presented in the following. The admission control strategy which basically ensures compliance with SLA for newly accepted as well as remaining users in the system is identical to the process described in [1].

User $i \in I_{sv}$ requests data rate by solving its individual optimization problem:

$$R_i(p_v) = \arg \max_{R_i^{\min} \le R \le R_i^{\max}} [U_i(R) - p_v R], \qquad (16)$$

where p_v is the price of a unit of data rate offered by the VNO v. Since function $U_i(R)$ is increasing and strictly concave for $R_i^{\min} \le R \le R_i^{\max}$,

$$R_i(p_v) = (1/k_i)\left([1/p_v]_{u_i^{\min}}^{u_i^{\max}} + k_i R_i^{\min} - u_i^{\min}\right). \qquad (17)$$

Figure 5 shows rate (17) versus price.

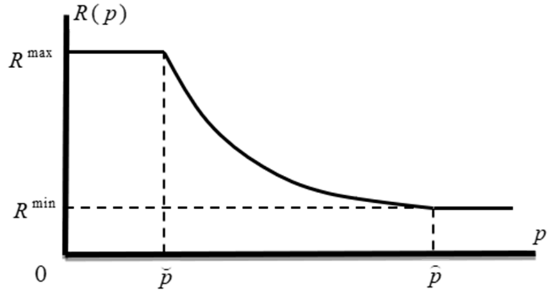

Figure 5. User rate vs. price

Due to the lower bound constraints in (7)-(8) optimization problem (5)-(8), (12)-(13) may not have a feasible solution. In this case, VNO v capacity deficit

$$\sum_{s=1}^{S} \sum_{i \in I_{sv}} R_i^{min} - C_v > 0 \;\; , v = 1,..,V \qquad (18)$$

is arbitrarily allocated to currently present users in this VNO.

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 $k = \{1,2,\ldots\}$. At each step k, users solve the individual optimization problems (16) resulting in rate (17). If constraints (7)-(8) are 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 in order to motivate users to request higher rate. However, if the constraints (7)-(8) are 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 capacity in an efficient way. The price adaptation model can be expressed as [16]:

$$p_{v,k+1} = \left[p_{v,k} + h(\tilde{R}_{vk} - C_{vk})\right]^{+} \qquad (19)$$

where

$$\tilde{R}_{vk} = max\left(C_v^{min}, \sum_{i \in I_{sv}} R_{ik}\right) \qquad (20)$$

$[x]^{+} = \max(0, x)$, 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.

In a slower time-scale each VNO adjusts its own capacity by negotiating the price with the VRRM. The adaptation of capacities among the tenant VNOs (C_v) is subject to the total available capacity of VRRM is C^{VRRM} (6). The average price of a unit of data rate in the entire system at step $k = \{1,2,\ldots\}$ is as follows:

$$P_k^{ave} = \frac{1}{C^{VRRM}} \sum_{v=1}^{V} C_v P_{v,k} \qquad (21)$$

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

We propose the following capacity adaptation algorithm for the VNOs according to [16]:

$$C_{v,k+1} = C_{v,k} + H(P_{v,k} - P_k^{ave}), \qquad (22)$$

where $H > 0$ is a small constant.

Algorithm (21)-(22) increases (decreases) the capacity of a VNO if its corresponding price is higher (lower) than the average price (21). However, VNO capacity cannot fall below the lower bound in (8) due to equation (20).

V. Simulation Scenario & Results

To evaluate our proposed resource management strategy, the simple traffic distribution scenario with VRRM capacity of 510 Mbps has been considered in this section. Network parameters are defined in Table 1. It is assumed that 3 VNOs with different SLA types (i.e., GB, BG and BE) are providing services from 4 service classes: *Conversational* (Con), *Streaming* (Str), *Interactive* (Int.) and *Background* (Bac.) according to the class-of-service definition in UMTS. 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 Internet of Things (IoT) and Email (Ema). It is further assumed that at each time step k, forty new users arrive and submit their requests for service to their associated VNOs. Simultaneously, twenty users depart from the system. For simplicity, the traffic type percentages of both arrivals and departures, defined as $U_{[\%]}^{srv}$, remain the same.

Table 1 – Network Parameters

VNO	**Service**	**Class**	$\boldsymbol{R_{svi}}$ in Mbps	$U_{[\%]}^{srv}$	$\boldsymbol{\lambda_s}$	C_v^{min} in Mbps
1 (GB)	Voi	Con.	[0.032, 0.064]	10	5	0.4 C^{VRRM}
	Vic		[1, 4]	10	4	
	Vis	Str.	[2, 13]	25	3	
	Mus		[0.064, 0.32]	15	1	
2 (BG)	Fil	Int.	[1, C^{VRRM}]	15	4	0.3 C^{VRRM}
	Web		[0.2, C^{VRRM}]	5	3	
	Soc		[0.4, C^{VRRM}]	10	2	
3 (BE)	Ema	Bac.	[0, C^{VRRM}]	5	4	0
	IoT		[0, 0.1]	5	4	

To evaluate the performance of user rate and VNO capacity adaptations, we consider stress scenario with proportionally increasing numbers of users of different services specified in Table 1. We assume user utility (2) as a particular case of utility-proportional rate allocation scheme. Benefits of utility-proportional fairness as compared to the previously used proportionally-fair strategy is demonstrated through extensive simulations. Figure 6 shows evolution of the system aggregate utility for utility proportionally rate allocation algorithm. The results show convergence of the users' rate adaptation and also highlights our assumption on separation of time scales, i.e., rate allocation should occurs much faster than changes in the set of users. The performance for the case where this assumption does not apply requires further investigation.

Figures 7 shows the converged system aggregate utility for utility proportional and proportionally fair resource allocation schemes. The advantage of utility-proportional scheme is clearly noticeable since admission of new users is inconsistent with decreasing aggregate utility. Figures 8 and 9 display another drawback of proportionally-fair rate allocation which gives advantage to users with lower rate SLA rate requirements. As the number of users increase at VNO GB, Voi users that have the lowest rate SLA requirements maintain their maximum requested rate of 0.064 Mbps long after Vis users rate drops to their minimum requested 2 Mbps. In general, under the proportionally fair resource allocation, users with highest SLA rate requirements will encounter reduced assigned rates well before users with lower SLA rates. Under a "fair" rate allocation scheme, all users should experience rate reduction from their maximum assigned rates approximately around the same time. As observed, this situation is better achieved using the utility-proportional strategy through proper adjustment of the parameter α in (2). Further results confirming this advantage have been omitted due to brevity.

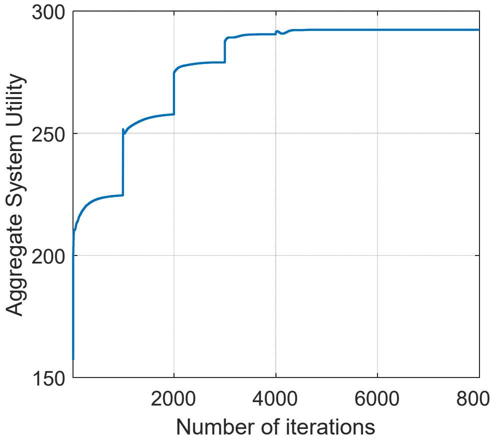

Figure 6. Evolution of aggregate utility for utility function (2)

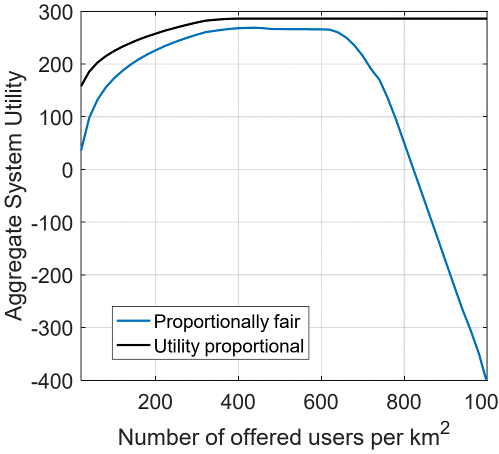

Figure 7. Aggregate utility for user utilities

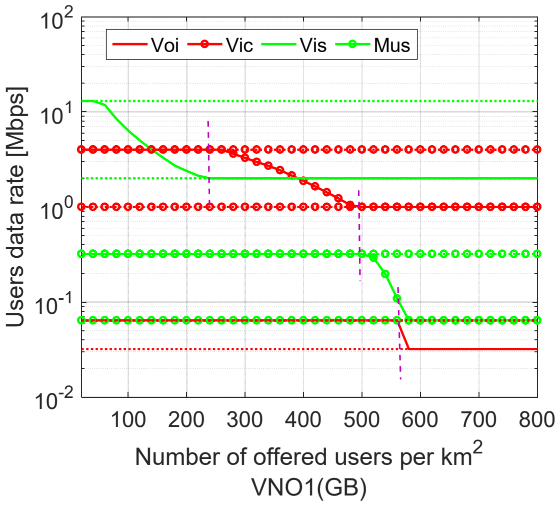

Figure 8. User rates for utilities (1)

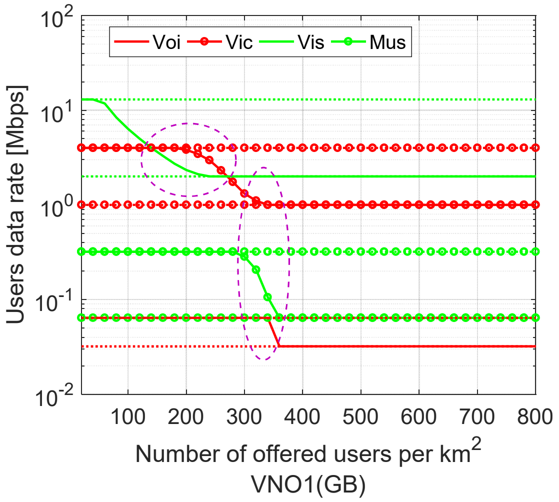

Figure 9. User rates for utilities (2)

VI. Conclusion and Future Research

The proposed radio resource management scheme in this research overcomes the shortcomings of the proportionally fair rate allocation by exploiting the general concept of utility proportional fairness. Simulation results clearly demonstrate advantages of using this strategy. In general, a customized SLA-based utility proportional fairness could lead to even better overall performance. The authors plan to further investigate this issue in future research. Viability of the time scale separation and convergence assumption in practical situations should also be studied. That will include mechanisms to mitigate performance loss in situations of comparable time scales in rate/capacity adaptation and users' arrivals/departures process. A requirement for this study is realistic models of users' arrival and departure processes. Finally, employing artificial intelligence (AI) techniques as a part of the network management could be a major focus of future research.

References

[1] B. Rouzbehani, V. Marbukh, and K. Sayrafian, "A Joint Admission Control & Resource Management Scheme for Virtualized Radio Access Networks", in *Proc. of CSCN'19 – 5th IEEE Conference on Standards for Communications and Networking*, Granada, Spain, Oct. 2019.

[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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] B. Rouzbehani, V. Marbukh, K. Sayrafian, and L.M. Correia, "Towards Cross-Layer Optimization of Virtualized Radio Access Networks," in *Proc. of EuCNC'19 – 28th European Conference on Networks and Communications,* Valencia, Spain, Jun. 2019.

[12] W.H. Wang, M. Palaniswami, and S. H. Low, "Application-oriented flow control: Fundamentals, algorithms and fairness," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1282 –1291, December 2006.

[13] 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.

[14] 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.

[15] C. Liu, L. Shi, and B. Liu, Utility-Based Bandwidth Allocation for Triple-Play Services, Fourth European Conference on Universal Multiservice Networks (ECUMN'07).

[16] 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, August 2006.