Performance Aware Resource Allocation and Traffic Aggregation for User Slices in Wireless HetNets

Georgios Smpokos[*†], Athanasios Lioumpas[*], Theodoros Mouroutis[*], Yiannis Stylianou[*] and Vangelis Angelakis[†]
[*]Department of Network Design, Research & Development, Cyta Hellas, Greece
Email: {georgios.sbokos, athanasios.lioumpas, theodoros.mouroutis, yiannis.stylianou}@hq.cyta.gr
[†]Department of Science & Technology, Linköping University, Sweden
Email: vangelis.angelakis@liu.se

***Abstract*—Resource allocation and traffic aggregation is becoming a significant part of the network deployment as it is expected that operators will offer advanced services through 5G systems. Scheduling procedures need to be fair but also to handle requirements set for diverse groups of users forming network slices. Heterogeneous networks will need to be controlled locally and the traffic could be split between the macro cell LTE nodes and small cell WiFi distributed access points (APs). In this work, we examine the performance of a novel optimal resource allocation algorithm while adding a new process aiming in achieving predefined throughput and delay performance for selected groups/slices of users. From the optimal resource allocation algorithm, we derive an algorithmic solution which can be applied to determine a number of network slices. The 3GPP-WLAN interworking setup evaluation shows that our proposal can be used for allocating resources to slices of users fulfilling throughput and delay requirements set by their subscription.**[*]

I. Introduction

The increasing demand for high data rates, low latency, energy efficiency and massive number of multipurpose devices connected to the Internet has already paved the way to a new era for wireless communications. The introduction of new standards and deployments will lead to the fifth generation (5G) of mobile network research and standardization. The upcoming surge in connectivity requirements (e.g. Internet of Things - IoT) and the increased demand for real time and steady performance services (e.g. video streaming) will stretch the already limited capacity and availability of the deployed networks [1], [2]. New physical layer methods that will increase spectral efficiency and system capacity have already been proposed including massive multiple-input-multiple-output (MIMO) antennas, full duplex and novel modulation schemes [3], [4].

Optimized resource utilization in heterogeneous networks (HetNets) is a research topic that will eventually lead to more efficient resource distribution and services implementing multiple technologies [4], [5], [6], [7]. Due to the fact that the current hardware based core and access network infrastructure cannot provide the necessary flexibility and efficient control of network components (switches, gateways, controllers), a new core-access backhaul network architecture needs to be adopted based on Software Defined Networking (SDN) and Network Function Virtualization (NFV) [3], [8], [9]. NFV and SDN will offer flexibility through the virtualization of the network components as well as the separation of the control and data plane (network subsystems control and user plane data flows). This in turn will offer flexible upgradability of components, optimization of traffic flow and customization of services [3], [10]. User group separation, also called slicing, will assist in providing dedicated services through the creation of virtual network slices that will reduce the signalling and control overhead, optimize the coordination of core with access RAN interfaces and subsystems and meet the specific requirements for groups of users [10], [11].

II. Motivation

In the context of offering multiple services, addressing diverse requirements (video, voice, file transfer, sensing) and connectivity through wireless HetNets to multipurpose devices (IoT, automotive, smartphones), mobile network operators (MNOs, MVNOs-virtual operators) should define and implement end to end slicing. Slicing will offer isolation, functional and performance independence and security in both core and access networks for the 5G ecosystem. Subscribers might be members of a number of slices based on their subscription agreements that will guarantee a minimum performance and service quality [10], [11], [12], [13]. To implement this, more network operators will install low power small cells (WiFi, LTE, LPWAN) offloading traffic from LTE and next generation macro cells. The deployment of small cells and traffic offloading could significantly improve the Quality of Service (QoS) and overall Quality of Experience (QoE) for the users in urban environments, while CAPEX and OPEX for operators and service providers will be reduced.

An architecture based on aggregating traffic is standardized in LTE Release 13 for cellular LTE and wireless LAN (WLAN) HetNets or LWA (LTE-WLAN Aggregation) [14] where the macro cell operates as an anchor and the small cell serves as a traffic booster. Following this trend, in [15] industrial researchers from Intel have recently proposed a process for splitting traffic in LTE-WiFi HetNets based on fairness while maximizing the total average UE throughput. Taking

[*]This work has been supported in part by the Horizon 2020 Marie Skłodowska-Curie Actions project WiVi-2020 (H2020-MSCA-ITN-2014-EID 642743-WiVi-2020).

all these into consideration Singh et al. [15] demonstrated the performance improvement for the average data rate of a group of users introducing a water-filling technique for allocating macro cell resources while users were aggregating traffic from small cells. The proposed low complexity implementation (optimal resource fraction algorithm) increased the throughput performance incorporating the aggregation principle while maximizing the proportional fairness. The portion of macro cell resources allocated to users was inversely proportional to the ratio of small cell data rate and macro cell spectral efficiency. The authors illustrated the performance gains while they set some simplified constraints e.g. constant backhaul delay for small cells and fixed traffic file size.

III. Contribution

The average delay values (latency) of a slice (group of subscribers) on a macro cell will need to fulfil the Key Performance Indicators (KPIs) defining that slice. Enhanced mobile broadband users (eMBB) will need higher average throughput availability compared to ultra-reliable low latency users (uRLL) that require low end to end latency [16]. In realistic scenarios the slices will be granted a limited portion of macro cell resources thus it is necessary to be able to control some parameters of the overall system such as the average throughput and average delay. The system will need to be adaptive as it could change rapidly by introducing new users, new slices or changes of resource availability at the macro cell.

Our goal is to enhance the concept described in [15] to control and select the parameters of the scheduling processes. This will allow us to fulfil specific slice/group KPIs related to throughput and end to end latency. In this work, we initially introduce a lower threshold than in [15] to compare the ratio of small cell downlink rate and macro cell peak capacity before allocating the macro cell resources portion to each user of that slice based on the optimal resource fraction algorithm. Deploying this algorithm we eventually exclude users with low throughput gains over their small cell connection and favour the allocation of macro cell resources to users with higher macro cell spectral efficiency. Adopting this strategy, the overall average delay of the group of users/slice is decreased while average throughput performance can fulfil the QoS requirements of the slice to which we allocate macro cell resources.

IV. System Model

We study a HetNet consisting an LTE macro cell (eNodeB) to which all users are registered and multiple small cells (e.g. WLAN APs) each user can be connected. Users that are out of the coverage area of any small cell are only connected to the macro cell. The small cells do not interfere in the frequency domain with the macro cell (different frequency bands). Traffic aggregation is based on traffic splitting at the macro cell node according to the anchor-booster framework where the LTE eNodeB acts as the anchor and the small cells as boosters [17]. Small cells are interconnected with the macro cell via a backhaul connection (fiber, mmWave, xDSL) for traffic transfer and signalling. Each UE k where $k = 1 \dots K$ and K is the number of users, has an instantaneous small cell data rate r_k known at macro cell and can aggregate traffic from the macro cell and small cell simultaneously. The delay of the backhaul connection l_k is technology and network topology specific and differs for each small cell.

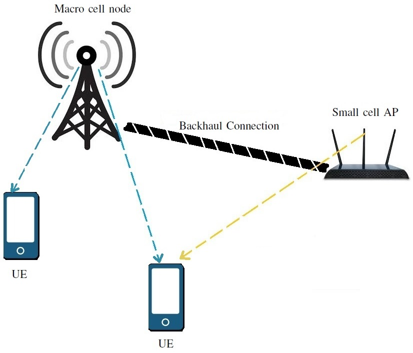

Fig. 1: Backhaul Connection between Macro cell and Small cell (control-data traffic) introducing delay in a traffic aggregation scenario (LWA).

Macro cell is aware of the spectral efficiency c_k of the LTE eNodeB-UE connection through CQI and CSI feedback mechanisms. The peak capacity p_k of that connection can be evaluated using the total available bandwidth W of the macro cell $p_k \triangleq c_k W$. This is the maximum macro cell data rate that each UE k can achieve connected to the macro cell when all the available resources of the macro cell are granted to it. Each UE k can be granted with a fraction of the macro cell resources denoted by $n_k \in [0, 1]$. The resources available to a group of users-slice can also be a fraction of the total amount of resources of the macro cell and can be expressed as n_s where $s = 1 \dots S$ is the slice index. An example of the system setup can be seen in Fig.1.

V. Problem formulation

Introducing the n_s slice resource fraction term we can observe how a variable portion of resources at the macro cell affects the performance of the system. In our study the backhaul delay l_k varies and is unique for each small cell-macro cell connection. The delay introduced by the backhaul interconnections adds on the total delay that each UE experiences. The total delay experienced by each user can be expressed as

$$l_{total,k} = l_{core,k} + l_k \tag{1}$$

where $l_{core,k}$ is the delay added by the core network.

For specific slices, it is necessary to reduce the total average delay but while maintaining an adequate throughput performance. Excluding small cell connections that do not contribute significantly to the UE's throughput performance can lead to an overall average delay reduction while the total average throughput is adequate for the slice requirements.

We assume there are K UEs forming a slice that can receive traffic from both the macro and small cell. We first try to maximize the logarithm of the sum of the rates for the UEs connected to the macro cell and a small cell as in [15]

$$\text{maximize} \sum_{k=1}^{K} \log(r_{eff,k} + n_k p_k) \tag{2a}$$

$$\text{subject to} \sum_{k=1}^{K} n_k = n_s \tag{2b}$$

where $r_{eff,k}$ is the small cell effective data rate for the k^{th} UE. This data rate is derived from the total time required for a UE to download a file with size f_k. Due to the delay l_k introduced by the backhaul, the total time for the k^{th} UE to download a file of size f_k is $t_k = l_k + \frac{f_k}{r_k}$. The effective data rate from the small cell for each user can be expressed as $r_{eff,k} \triangleq \frac{f_k}{t_k} = \left(\frac{1}{r_k} + \frac{l_k}{f_k}\right)^{-1}$. For UEs that are not inside the range of any small cell $r_k = r_{eff,k} = 0$. The fraction $n_s \in [0,1], \sum_{s=1}^{S} n_s = 1$ indicates the portion of resources of the macro cell allocated to a slice of users.

In this work, we will exploit a mechanism that excludes (deactivates) low throughput small cell connections before assigning to the UEs the macro cell portion of resources. This will reduce the total average delay by not aggregating traffic from small cells that do not contribute significantly in terms of throughput compared to the macro cell peak capacity. This problem will eventually become a tradeoff between average throughput and average delay performance for a group of UEs/slice.

VI. Algorithmic Solution

In order to solve the problem expressed in (2a)-(2b) we will use the method of Lagrange multipliers where a Lagrange multiplier ν is introduced. Expression (2a) then becomes

$$\text{maximize} \sum_{k=1}^{K} \log(r_{eff,k} + n_k p_k) + \nu \left(\sum_{k=1}^{K} n_k - n_s \right) \tag{3}$$

After we differentiate (3) with respect to n_k $(\partial/\partial n_k)$ and setting equal to zero

$$\frac{p_k}{r_{eff,k} + n_k p_k} = -\nu, \quad \forall k \in [1, K] \tag{4}$$

we can further simplify (4)

$$\frac{r_{eff,k}}{p_k} + n_k = \mathrm{A}, \quad \forall k \in [1, K] \tag{5}$$

where $\mathrm{A} = -1/\nu$ is a constant to meet the resource allocation constraint. Calculating the sum of (5) for all of the UEs we have

$$\frac{1}{K} \left(\sum_{k=1}^{K} \frac{r_{eff,k}}{p_k} + n_s \right) = \mathrm{A} \ . \tag{6}$$

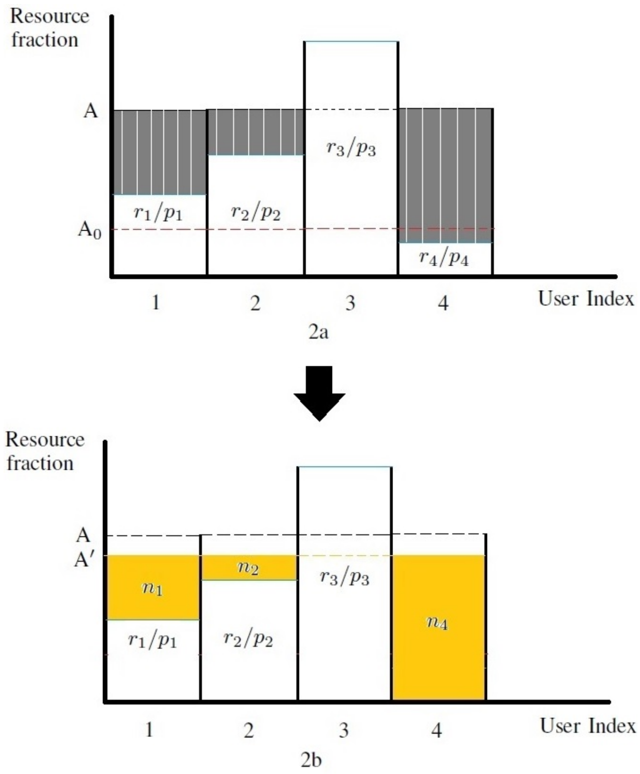

Fig. 2: Algorithm steps based on Water-Filling technique

In order to discard small cell connections that do not contribute significantly to the total aggregated UE throughput we introduce another lower limit A_0 based on A

$$\mathrm{A}_0 = \frac{\mathrm{A}}{\lambda}, \quad \lambda \in \mathbb{R}^+ \tag{7}$$

The denominator λ is a number selected to set a lower limit A_0 that is a portion of A. The algorithm first compares the rate ratio $\frac{r_{eff,k}}{p_k}$ with the lower limit A_0 and if it is lower disables the small cell connection resulting in $r_k = r_{eff,k} = 0$. By doing this the backhaul delay component is eliminated $l_k = 0$. After the rate ratios are updated the algorithm calculates a new level A' using (6). Next, if required $\big($rate ratio $\frac{r_{eff,k}}{p_k}$ higher than updated constant level $\mathrm{A}'\big)$ the algorithm eliminates the macro cell resource fraction allocation for UE k ($n_k = 0$) as these users gain less from aggregating traffic from the macro cell. In Fig. 2 we can observe a representation of the algorithm allocation mechanism. The rate ratio values for each user are illustrated by the height of the bars.

Example: Let us consider the case where there are four users $K = 4$ (Fig. 2) forming a single slice with reduced delay requirements and half of the macro cell resources assigned to it ($n_s = 0.5$). The macro cell and small cells update the information necessary to calculate the peak capacity p_k and the small cell effective rate $r_{eff,k}$. The algorithm then calculates "sea" level A. After we introduce a term λ we calculate the ratio $\mathrm{A}_0 = \frac{\mathrm{A}}{\lambda}$ to set a lower limit for the $\frac{r_{eff,k}}{p_k}$ comparison. As we observe from Fig. 2a user's 4 rate ratio is below A_0

threshold. User 4 will not aggregate traffic from the small cell as there is no significant throughput gain. After setting $r_{eff,4} = 0$ the algorithm calculates the new upper threshold A$'$ which is lower than the initial A. Next, the algorithm checks if the rate ratio $\frac{r_{eff,k}}{p_k}$ is above threshold A$'$ that happens for user 3 $\left(\frac{r_{eff,3}}{p_3} \geq \mathrm{A}'\right)$. User 3 will not aggregate traffic from the macro cell as the small cell data rate is adequate or its peak capacity (macro cell connection) is not significant.

Algorithm 1 Slice resource fraction algorithm

```
procedure SLICE-ALLOC
    $n = 1, B = \mathrm{A}$
    while $\frac{r_{eff,n}}{p_n} \leq B/\lambda$ do
        $r_{eff,n} = 0$
        $n = n + 1$
    end while
    $N = K,\ B = \frac{1}{N}\left(\sum_{n=1}^{N} \frac{r_{eff,n}}{p_n} + n_s\right)$
    sort indices such that $\frac{r_{eff,1}}{p_1} \leq \frac{r_{eff,2}}{p_2} \leq \ldots \frac{r_{eff,K}}{p_K}$
    while $n_N = B - \frac{r_{eff,N}}{p_N} \leq 0$ do
        $N = N - 1$
        $B = \frac{1}{N}\left(\sum_{n=1}^{N} \frac{r_{eff,n}}{p_n} + n_s\right)$
    end while
    $n_k = B - \frac{r_{eff,k}}{p_k}\ \forall k = 1 \ldots N;\ n_k = 0\ \forall k = N + 1 \ldots K$
    Unsort user indices
end procedure
```

VII. Performance evaluation

The proposed slice resource fraction algorithm is compared with the optimal allocation algorithm of [15]. In our simulations both algorithms are evaluated in terms of average throughput and delay performance while the macro cell RAT is LTE and the small cells are IEEE 802.11n WLAN APs. The backhaul delay l_k introduced by the macro cell-small cell connection is selected randomly following a uniform distribution. User slices, as mentioned before, can have different requirements, thus it is beneficial for us to explore how the selection of the denominator constant λ affects the performance of the system in terms of average throughput and average delay, taking into consideration the total number of users per sector K and resource portion (n_s) availability.

In our simulations we use [15] as a reference in order to demonstrate the benefits of the proposed algorithm in terms of throughput-delay tradeoff manipulation. In order to do that we define a ratio for both the average throughput and average delay as $R_T = T_{avg,SLICE-ALLOC}/T_{avg,OPT-ALLOC}$ and $R_D = D_{avg,SLICE-ALLOC}/D_{avg,OPT-ALLOC}$ respectively comparing the average throughput and delay of the proposed algorithm and the one in [15]. We also introduce another measure we call throughput-delay difference $\delta = R_T - R_D$ which gives us an indication of the margin comparing the throughput and delay reduction. It is of our interest to maximize that margin, thus experiencing a small reduction in average throughput but instantly a higher reduction in average delay leading to a total reduction in latency. Practically, this can be translated into an operator choosing the sweet spot for the resource shceduling and aggregation deciding the tradeoff betweeen the average throughput and the delay of the serviced slice.

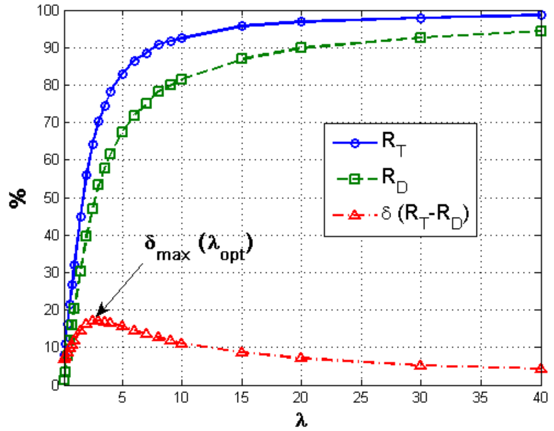

Fig. 3: Difference δ $(R_T - R_D)$ vs. λ values for 30 users and $n_s = 1$.

In Fig. 3 we can see the reduction in percentage (%) of the total average throughput and total average delay for different values of λ. The simulation was set for 30 users per sector and all the resources of that sector were available to all users ($n_s = 1$, i.e. single slice). As λ increases ($\lambda \to \infty$) the proposed algorithm behaves as the optimal allocation algorithm of [15]. For small enough values of λ there is a reduction in average throughput and average delay of the system. The difference in percentage of throughput and delay δ is also illustrated in Fig. 3 and exhibits a maximum. An operator may select to have that difference maximized resulting in reducing significantly the average delay (approximatelly 50%) while experiencing low reduction in average throughput.

The optimum λ (δ_{max}) value is only related to the total number of users K of the slice and the portion of resources (slice portion) dedicated to that set of users (n_s). The relationship between number of users, slice portion, λ optimum and maximum difference δ_{max} value can be seen in Fig. 4 and Fig. 5. A controller stationed at the macro cell could select an appropriate λ value in order to fulfil the requirements of the set of users of the slice (allocate macro cell resources and aggregate traffic from the small cell). Selecting the appropriate thresholds using our slice resource fraction allocation algorithm we control the overall throughput and delay performance. Knowing (reporting, statistical analysis) the maximum capacity and throughput performance of the system (macro cell and small cells) and the average backhaul delay, the operator can select the parameter λ for the slice resource fraction algorithm in order to set the average delay reduction and throughput levels.

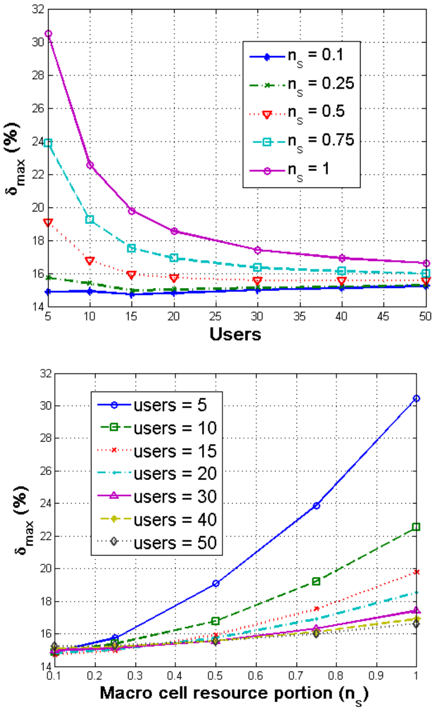

Fig. 4: Maximum difference δ_{max} vs. number of users and macro cell portion n_s.

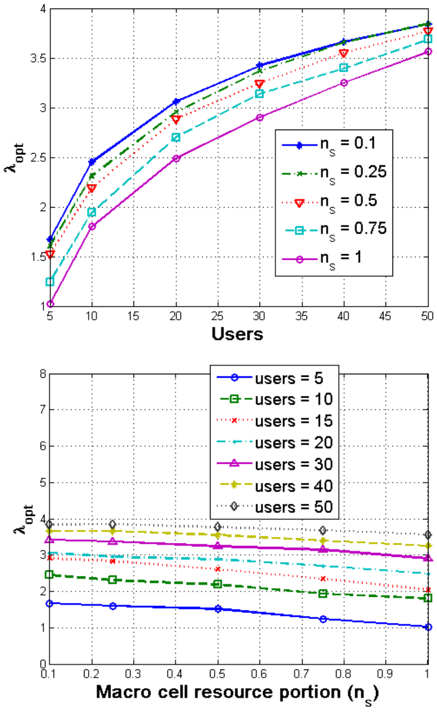

Fig. 5: λ_{opt} (δ_{max}) vs. number of users and macro cell portion n_s.

VIII. Conclusion

In this work, we demonstrated how an evolved form of the LWA optimal resource allocation algorithm introduced by the authors of [15] can be deployed. Our proposal aims to modify appropriately the throughput and delay performance of a group of users served by a macro cell and distributed WiFi small cells. The evaluation of the proposed solution showed that by selecting appropriate values for the slice resource fraction algorithm we can set the overall performance of a group of users without affecting the performance of other user groups/slices within the same macro cell sector. This process can be mainly utilized in an offline evaluation setup as it is computationally complex. It could be beneficial then for later work to investigate an online fast converging procedure to select the λ values based on this process or find closed form formulas. The λ selection will let us achieve specific performance requirements for slices served continuously within the 3GPP-LWA framework.

References

[1] EU Digital Economy and Society Commission, *5G Manifesto for timely deployment of 5G in Europe*. 5G Manifesto, July 2016.

[2] Cisco, *Cisco visual networking index: Global mobile data traffic forecast update 2014-2019*. White Paper, February 2015.

[3] M. Jaber, M. Imran, R. Tafazolli, and A. Tukmanov, *5G Backhaul Challenges and Emerging Research Directions: A Survey*, 2016, Volume: 4, Pages: 1743 - 1766, DOI: 10.1109/ACCESS.2016.2556011.

[4] C. X. Wan et al, *Cellular Architecture and Key Technologies for 5G Wireless Communication Networks*, IEEE Communications Magazine, February 2014, Volume: 52, Issue: 2, Pages: 122 - 130, DOI: 10.1109/MCOM.2014.6736752.

[5] M. Gerasimenko et al, *Cooperative Radio Resource Management in Heterogeneous Cloud Radio Access Networks*, IEEE Access, 2015, Volume: 3, Pages: 397 - 406, DOI: 10.1109/ACCESS.2015.2422266.

[6] S. Borst, S. Hanly, and P. Whiting, *Optimal Resource Allocation in HetNets*, IEEE ICC 2013 - Wireless Communications Symposium, 2013, Pages: 5437 - 5441, DOI: 10.1109/ICC.2013.6655454.

[7] V. Angelakis et al, *Allocation of Heterogeneous Resources of an IoT Device to Flexible Services*, IEEE Internet of Things Journal, 2016, Volume: 3, Issue: 5.

[8] J. Zhang, W. Xie, and F. Yiang, *An Architecture for 5G Mobile Network Based on SDN and NFV*, ICWMMN2015 Proceedings, 2015, Pages: 87 - 92, DOI: 10.1049/cp.2015.0918.

[9] Q. Zhou, C. X. Wang, S. McLaughlin, and X. Zhou, *Network Virtualization and Resource Description in Software-Defined Wireless Networks*, IEEE Communications Magazine, November 2015, Volume: 53, Issue: 11 Pages: 110 - 117, DOI: 10.1109/MCOM.2015.7321979.

[10] X. An et al, *On end to end network slicing for 5G communication systems*, Transactions on Emerging Telecommunications Technologies, Wiley Online Library, 2016.

[11] V. G. Nguen and Y. H. Kim, *Slicing the Next Mobile Packet Core Network*, 2014 11th International Symposium on Wireless Communications Systems (ISWCS), 2014, Pages: 901 - 904, DOI: 10.1109/ISWCS.2014.6933481.

[12] C. Liang and F. R. Yu, *Wireless Virtualization for Next Generation Mobile Cellular Networks*, IEEE Wireless Communications, 2015, Volume: 22, Issue: 1 Pages: 61 - 69, DOI: 10.1109/MWC.2015.7054720.

[13] R. Riggio et al, *Scheduling Wireless Virtual Networks Functions*, IEEE Transactions on Network and Service Management, June 2016, Volume: 13, Issue: 2, Pages: 240 - 252, DOI: 10.1109/TNSM.2016.2549563.

[14] 3GPP, *Introduction of LTE-WLAN Radio Level Integration and Interworking Enhancement stage-2,*, Tech. Rep. R2-156737, Nov. 2015.

[15] S. Singh et al, *Proportional Fair Traffic Splitting and Aggregation in Heterogeneous Wireless Networks*, IEEE Communications Letters March 2016, Volume: 20, Issue: 5 Pages: 1010 - 1013, DOI: 10.1109/LCOMM.2016.2547418.

[16] NGMN Alliance, *NGMN 5G White Paper,*, Feb. 2015.

[17] A. Zakrzewska et al, *Dual connectivity in LTE HetNets with split control and user plane,* IEEE Globecom Workshops, pp. 391-396, Dec. 2013.