Delay-aware and power-efficient resource allocation in virtualized wireless networks

Delay-aware and Power-Efficient Resource Allocation in Virtualized Wireless Networks

Saeedeh Parsaeefard*, Vikas Jumba†, Mahsa Derakhshani‡, Tho Le-Ngoc†
* Iran Telecommunication Research Center, Tehran, Iran
†Department of Electrical & Computer Engineering, McGill University, Montreal, QC, Canada
‡Department of Electrical & Electronic Engineering , Imperial College, London, UK

***Abstract*—This paper proposes a delay-aware resource provisioning policy for virtualized wireless networks (VWNs) to minimize the total average transmit power while holding the minimum required average rate of each slice and maximum average packet transmission delay for each user. The proposed cross-layer optimization problem is inherently non-convex and has high computational complexity. To develop an efficient solution, we first transform cross-layer dependent constraints into physical layer dependent ones. Afterwards, we apply different convexification techniques based on variable transformations and relaxations, and propose an iterative algorithm to reach the optimal solution. Simulation results illustrate the effects of the required average packet transmission delay and minimum average slice rate on the total transmission power in VWN.**

***Index Terms*—Delay-aware resource provisioning, virtualized wireless networks.**

I. Introduction

Wireless network virtualization has been recently considered as a promising approach to enhance spectrum efficiency via sharing infrastructures among different service providers (also called slices) serving their own specific sets of users [1]. Successful sharing requires proper isolation between slices to prevent harmful effects of user activity in one slice on the other slices [2].

Various resource provisioning policies in VWNs are being proposed to provide this isolation either by static resource allocation to each slice or dynamic throughput reservation [3]. In [4], authors propose a Karnaugh-map-like online embedding algorithm for VWN to handle network requests. In [5], the concept of game theory is used to provide dynamic interactions between slices and network operators. An opportunistic spectrum sharing method in VWNs with multiple physical networks is proposed in [6]. The resource allocation schemes in VWNs to maximize the total rate under various quality-of-service (QoS) requirements of slices are investigated in [7], [8] for OFDMA and massive MIMO based VWNs, respectively.

While the above works focus on the physical-layer parameters to provide the isolation among slices, it is essential to consider the traffic characteristics of end-users, e.g., arrival rate and tolerable delay, to ensure high-quality end-user experience in VWN. This issue calls for a new constraint related to maximum delay of each user in resource allocation problem. In addition, energy efficiency should be considered as an important objective of VWNs for the next generation of wireless networks [2]. In this paper, we propose cross-layer resource provisioning policies suitable for VWNs to minimize the power consumption in consideration of traffic arrival rate and tolerable delay while satisfying the dynamic slice isolation.

We consider an up-link transmission of an orthogonal frequency division multiple access (OFDMA) based VWN. In this setup, we propose a delay-aware resource allocation that minimizes the total average power of VWN while holding the minimum average rate of each slice for isolation and limiting the packet transmission to a maximum delay. Due to these two constraints, the formulated resource allocation problem has an inherent cross-layer as well as non-convex nature and suffers from high computational complexity. To reach a tractable formulation and capture the meaning of tolerable delay from end-users perspectives, we resort to the concept of effective capacity [9]–[11]. By replacing the delay-aware constraints with more tractable formulations and applying relaxation and transformation techniques, we convexify the formulated problem and develop an efficient iterative algorithm for solution. Through simulations, we investigate the effects of different parameters on the transmit power of VWN for two user-location scenarios: cell-center and cell-boundary. We show how packet size, arrival rate of users and maximum tolerable delay can affect the total transmit power of VWN.

The rest of this paper is organized as follows. Section II introduces the system model and problem formulation. Section III develops the proposed iterative algorithm for slice provisioning. Section IV presents the simulation results and Section V concludes the paper.

II. System Model

We consider an up-link OFDMA transmission of a single-cell VWN, where the base station (BS) is virtualized to support a set of $\mathcal{G} = \{1, \cdots, G\}$ slices. Each slice $g \in \mathcal{G}$ serves a set of $\mathcal{N}_g = \{1, \cdots, N_g\}$ users and $N = \sum_{g \in \mathcal{G}} N_g$ is the total number of users in the VWN. The total wireless channel bandwidth B is equally divided into $\mathcal{K} = \{1, \cdots, K\}$ set of OFDMA sub-carriers. The bandwidth $B_k = B/K$ of each sub-carrier $k \in \mathcal{K}$ is assumed to be less than the coherence bandwidth B_c. Therefore, the link from user n_g to the BS on sub-carrier k exhibits flat fading with channel power gain $h_{n_g,k}$. It is also assumed that the overall channel power gain vector $\mathbf{h} = [h_{n_g,k}]_{\forall n_g, \forall g, \forall k}$ has a known stationary and ergodic cumulative distribution function (cdf) [12], [13].

Let $w_{n_g,k}$ denote the sub-carrier assignment indicator, where $w_{n_g,k} = 1$ indicates that the sub-carrier k is assigned to user n_g, otherwise $w_{n_g,k} = 0$. With this notation, the OFDMA exclusive sub-carrier assignment policy can be expressed as the following constraint

$$\text{C1: } \sum_{g\in\mathcal{G}} \sum_{n_g\in\mathcal{N}_g} w_{n_g,k} \leq 1 \text{ and } w_{n_g,k} \in \{0,1\},\ \forall k \in \mathcal{K}.$$

If $p_{n_g,k}$ is the power allocated to user n_g on sub-carrier k in time slot t, the corresponding achievable rate is $R_{n_g,k} = \log_2(1 + \frac{p_{n_g,k}h_{n_g,k}}{\sigma})$, where σ is the noise power in each sub-carrier and users. Consequently, the total achievable rate of user n_g becomes $R_{n_g}(\mathbf{P},\mathbf{W}) = \sum_{k\in\mathcal{K}} R_{n_g,k}$, where $\mathbf{P} = [p_{n_g,k}]_{\forall n_g,\forall g,\forall k}$ and $\mathbf{W} = [w_{n_g,k}]_{\forall n_g,\forall g,\forall k}$ are the allocated power and sub-carrier vector to all users of all slices. To maintain isolation among slices and offer reliable QoS to each user in all slices, we consider following constraints:

- ***Slice-Isolation constraints:*** by guaranteeing the minimum average rate R_g^{rsv} i.e.,

$$\text{C2: } \mathbf{E_h}\left\{\sum_{n_g\in\mathcal{N}_g} R_{n_g}(\mathbf{P},\mathbf{W})\right\} \geq R_g^{\text{rsv}},\ \forall g \in \mathcal{G},$$

  where operator $\mathbf{E_h}\{\cdot\}$ represents the expectation over random vector $\mathbf{h}$.
- ***User QoS constraints:*** by keeping the average traffic delay for each user n_g for all $g \in \mathcal{G}$ below the predefined threshold, α_{n_g} (Eq. 17 in [14]), i.e.,

$$\text{C3: } \alpha_{n_g}\mathbf{E_h}\{Q_{n_g}\} \leq D_{n_g},\ \forall n_g \in \mathcal{N}_g,\ \forall g \in \mathcal{G},$$

  where Q_{n_g} and α_{n_g} denote the queue length and average packet arrival rate at the queue of user n_g, respectively.

With considering energy efficiency for VWN under the above two types of constraints, the overall optimization problem can be written as

$$\min_{\mathbf{P},\mathbf{W}}\ \mathbf{E_h}\left\{\sum_{g\in\mathcal{G}}\sum_{n_g\in\mathcal{N}_g}\sum_{k\in\mathcal{K}} w_{n_g,k}p_{n_g,k}\right\} \quad (1)$$

subject to C1-C3.

Note that the proposed resource provisioning problem (1) contains discrete and continuous variables such as $\mathbf{W}$ and $\mathbf{P}$, and is inherently non-convex in nature [15]. Additionally, the constraints from both physical layer (C1 and C2) and medium access control (MAC) layer (C3) add multi dimensional complexity to the optimization problem.

III. Proposed Algorithm

C3 contributes to the major complexity of (1) due to its relationship with cross-layer parameters. As a first step to simplify (1), we need to find the equivalence of C3 in terms of $\mathbf{P}$ and $\mathbf{W}$ instead of Q_{n_g} and α_{n_g}. In this context, according to Lemma 1 in [9], C3 can be rewritten as

$$\widehat{\text{C3}}:\ \mathbf{E_h}\left[R_{n_g}(\mathbf{P},\mathbf{W})\right] \geq Z_{n_g},\ \forall n_g \in \mathcal{N}_g,\ \forall g \in \mathcal{G},$$

where

$$Z_{n_g} = \frac{(2D_{n_g}\alpha_{n_g}+2) + \sqrt{(2D_{n_g}\alpha_{n_g}+2)^2 - 8D_{n_g}\alpha_{n_g}}}{4D_{n_g}} L_{n_g}$$

and L_{n_g} is the average packet size at the queue of user n_g. In the next step, to get rid of combinatorial structure of problem, we apply the relaxation and re-transformation techniques to (1) and consider $w_{n_g,k} \in [0,\ 1]$, representing the sub-carrier assignment for the fraction of a time slot [16]. Based on new transformations, C1 is changed to

$$\widetilde{\text{C1}}:\ \sum_{g\in\mathcal{G}}\sum_{n_g\in\mathcal{N}_g} w_{n_g,k} \leq 1 \text{ and } w_{n_g,k} \in [0,\ 1], \forall k \in \mathcal{K}.$$

Furthermore, in order to convexify (1), we consider a new variable $x_{n_g,k} = p_{n_g,k}w_{n_g,k}$ to transform the rate as

$$\widetilde{R}_{n_g}(\mathbf{X},\mathbf{W}) = \sum_{k\in\mathcal{K}} w_{n_g,k}\log_2(1 + \frac{x_{n_g,k}h_{n_g,k}}{\sigma w_{n_g,k}}),$$

where $\mathbf{X}$ is the vector of $x_{n_g,k}$ for all users and sub-carriers. The above transformed throughput belongs to the general class of convex function $f(x,y) = x\log(1+\frac{y}{x})$ [15], [17]. Therefore, C2 and C3 become

$$\widetilde{\text{C2}}:\ \mathbf{E_h}\left\{\sum_{n_g\in\mathcal{N}_g}\widetilde{R}_{n_g}(\mathbf{X},\mathbf{W})\right\} \geq R_g^{\text{rsv}},\ \forall g \in \mathcal{G} \text{ and}$$

$$\widetilde{\text{C3}}:\ \mathbf{E_h}\left[\widetilde{R}_{n_g}(\mathbf{X},\mathbf{W})\right] \geq Z_{n_g},\ \forall n_g \in \mathcal{N}_g,\ \forall g \in \mathcal{G},$$

respectively. Consequently, (1) can be rewritten as

$$\min_{\mathbf{X},\mathbf{W}}\ \mathbf{E_h}\left\{\sum_{g\in\mathcal{G}}\sum_{n_g\in\mathcal{N}_g}\sum_{k\in\mathcal{K}} x_{n_g,k}\right\} \quad (2)$$

subject to $\widetilde{\text{C1}} - \widetilde{\text{C3}}$

Now, since (2) is a convex problem, it can be solved via Lagrange dual method [18], [19] and the optimal solution can be achieved for this scenario. The Lagrange function of (2) is

$$\begin{aligned}\mathcal{L}(\phi_g,\zeta_{n_g},\rho_k,\mathbf{X},\mathbf{W}) = &-\mathbf{E_h}\left\{\sum_{g\in\mathcal{G}}\sum_{n_g\in\mathcal{N}_g}\sum_{k\in\mathcal{K}} x_{n_g,k}\right\} \quad (3)\\ &+\sum_{g\in\mathcal{G}}\phi_g(\mathbf{E_h}\left\{\sum_{n_g\in\mathcal{N}_g}\widetilde{R}_{n_g}\right\} - R_g^{\text{rsv}})\\ &+\sum_{g\in\mathcal{G}}\sum_{n_g\in\mathcal{N}_g}\zeta_{n_g}(\mathbf{E_h}\left\{\widetilde{R}_{n_g}\right\} - Z_{n_g})\\ &+\sum_{k\in\mathcal{K}}\rho_k\left(1 - \sum_{g\in\mathcal{G}}\sum_{n_g\in\mathcal{N}_g} w_{n_g,k}\right),\end{aligned}$$

where ρ_k for all $k \in \mathcal{K}$, ϕ_g for all $g \in \mathcal{G}$ and ζ_{n_g} for all $n_g \in \mathcal{N}_g$ are the positive Lagrange variables of $\widetilde{\text{C1}}$, $\widetilde{\text{C2}}$ and $\widetilde{\text{C3}}$, respectively. Let $\boldsymbol{\rho}$, $\boldsymbol{\phi}$ and $\boldsymbol{\zeta}$ be the vectors of the corresponding Lagrange variables ρ_k, ϕ_g and ζ_{n_g}, respectively.

Now, the Dual function related to (3) is

$$\mathcal{D}(\boldsymbol{\phi},\boldsymbol{\zeta},\boldsymbol{\rho}) = \max_{\mathbf{X},\mathbf{W}} \mathcal{L}(\boldsymbol{\phi},\boldsymbol{\zeta},\boldsymbol{\rho},\mathbf{X},\mathbf{W})$$

and consequently, the dual problem is

$$\min_{\boldsymbol{\phi},\boldsymbol{\zeta},\boldsymbol{\rho}} \ \mathcal{D}(\boldsymbol{\phi},\boldsymbol{\zeta},\boldsymbol{\rho}) \quad \text{subject to } \widetilde{\text{C1}} - \widetilde{\text{C3}}. \tag{4}$$

For this type of resource allocation problem, the duality gap is zero for large number of sub-carriers, i.e., the solution of dual problem is equivalent to the solution of primal problem [13], [20]. By applying the KKT condition to (4), the optimal power for user n_g on sub-carrier k, $p^*_{n_g,k}$, is

$$p^*_{n_g,k} = \left[\frac{\phi_g + \zeta_{n_g}}{\ln(2)} - \frac{\sigma}{h_{n_g,k}}\right]_0^{p_{\max}}, \tag{5}$$

where $[x]_b^a = \max\{\min\{x,a\},b\}$. From KKT conditions, for optimal sub-carrier allocation $w^*_{n_g,k}$, we have,

$$w^*_{n_g,k} \begin{cases} = 0, & \frac{\partial \mathcal{L}(\phi_g,\zeta_{n_g},\rho_k,\mathbf{X},\mathbf{W})}{\partial w^*_{n_g,k}} < 0 \\ \in [0,1], & \frac{\partial \mathcal{L}(\phi_g,\zeta_{n_g},\rho_k,\mathbf{X},\mathbf{W})}{\partial w^*_{n_g,k}} = 0 \\ = 1, & \frac{\partial \mathcal{L}(\phi_g,\zeta_{ng},\rho_k,\mathbf{X},\mathbf{W})}{\partial w^*_{n_g,k}} > 0, \end{cases}$$

where

$$\frac{\partial \mathcal{L}(\phi_g,\zeta_{n_g},\rho_k,\mathbf{X},\mathbf{W})}{\partial w^*_{n_g,k}} = (\phi_g + \zeta_{n_g})\times \left(\log_2(1+\gamma_{n_g,k}) - \frac{\gamma_{n_g,k}}{(1+\gamma_{n_g,k})\ln(2)}\right)$$

and $\gamma_{n_g,k} = \frac{x_{n_g,k}h_{n_g,k}}{\sigma w_{n_g,k}}$. In order to hold the exclusive sub-carrier allocation of OFDMA, the sub-carrier k is allocated to user which satisfy the followings

$$w^*_{n'_g,k} = \begin{cases} 1, & n'_g = \max_{\forall n_g \in \mathcal{N}_g, \forall g \in \mathcal{G}} \frac{\partial \mathcal{L}(\phi_g,\zeta_{n_g},\rho_k,\mathbf{X},\mathbf{W})}{\partial w^*_{n_g,k}}, \\ 0, & n_g \neq n'_g. \end{cases} \tag{6}$$

The iterative algorithm to derive the optimal sub-carrier and power algorithm for (2), summarized in Algorithm 1, consists of two phases:

1) Off-line Phase: Generate different channel state information (CSI) samples of VWN according to a known channel distribution information (CDI) of users. Via these samples and gradient descent method, the Lagrange variables $\boldsymbol{\phi}$ and $\boldsymbol{\zeta}$ are updated as

$$\phi^*_g(j) = \phi_g(j-1) + \delta_{\phi_g}\left(\frac{\partial \mathcal{L}}{\partial \phi_g}\right),\ \forall g \in \mathcal{G} \text{ and}$$

$$\zeta^*_{n_g}(j) = \zeta_{n_g}(j-1) + \delta_{\zeta_{n_g}}\left(\frac{\partial \mathcal{L}}{\partial \zeta_{n_g}}\right),\ \forall n_g \in \mathcal{N}_g,\ \forall g \in \mathcal{G},$$

where $0 < \delta_{\phi_g} \ll 1$ and $0 < \delta_{\zeta_{n_g}} \ll 1$ are small positive step sizes for ϕ_g and ζ_{n_g}, respectively. The update functions can be equivalently rewritten as

$$\phi^*_g(j) = \phi_g(j-1) + \delta_{\phi_g}\left(\mathbf{E_h}\left\{\sum_{n_g \in \mathcal{N}_g} \widetilde{R}_{n_g}(\mathbf{X}(j-1),\mathbf{W}(j-1))\right\} - R_g^{\text{rsv}}\right), \tag{7}$$

$$\zeta^*_{n_g}(j) = \zeta_{n_g}(j-1) + \delta_{\zeta_{n_g}}\left(\mathbf{E_h}\left\{\widetilde{R}_{n_g}(\mathbf{X}(j-1),\mathbf{W}(j-1))\right\} - Z_{n_g}\right). \tag{8}$$

The iterative process will be terminated if

$$\|\phi^*_g(j) - \phi_g(j-1)\| \leq \varepsilon_1 \text{ and } \|\zeta^*_{n_g}(j) - \zeta_{n_g}(j-1)\| \leq \varepsilon_2 \tag{9}$$

where $0 < \varepsilon_1 \ll 1$ and $0 < \varepsilon_2 \ll 1$.

2) On-line Phase: Power and sub-carrier are allocated according to (5) and (6) at each time slot, based on the values derived from off-line phase for fixed CDIs.

Algorithm 1 : Slice Provisioning Algorithm

Off-line Phase:

Initialization: Set arbitrary values for $j = 0$, $\boldsymbol{\phi}(j=0)$, $\boldsymbol{\zeta}(j=0)$, $\delta_{\phi_g}(j=0)$, $\delta_{\zeta_{n_g}}(j=0)$, and $j_{\max} \gg 1$, as the maximum number of iteration for offline phase.

Repeat: $j = j+1$

Update $\boldsymbol{\phi(j)}$ and $\boldsymbol{\zeta(j)}$ according to (7) and (8).

Until (9) holds or $j \geq j_{\max}$.

On-line Phase:

Update **P** and **W** according to (5) and (6), respectively.

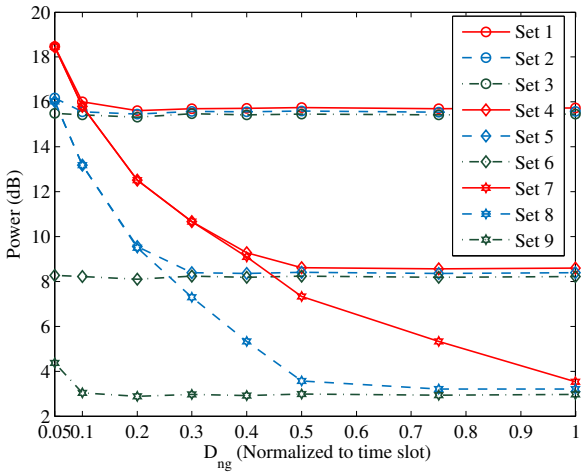
If CDI changes, go to **Off-line Phase**, otherwise, continue **On-line Phase** for next time instance.

When CDI is changed, **Off-line Phase** is executed to up-date $\boldsymbol{\phi}$ and $\boldsymbol{\zeta}$ [12], [13].

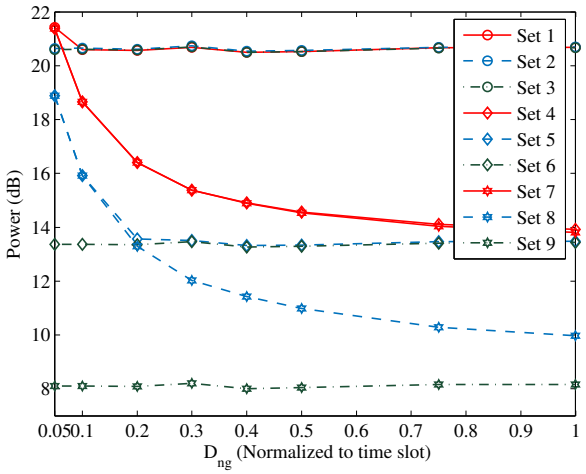
IV. Simulations Results

In this section, we evaluate the performance of proposed algorithm via simulation results. We consider a single BS serving $G = 2$ slices with $K = 64$ OFDMA sub-carriers. VWN has $N = 4$, total number of users with each slice having $N_1 = N_2 = 2$ users. Each slot length t is normalized to one. Furthermore, we set maximum average delay of packet transmission at each user's queue $D_{n_g} = 0.5$, minimum reserved rate $R^{\text{rsv}}_{g_1} = R^{\text{rsv}}_{g_2} = 1.0$ bps/Hz for slices g_1 and g_2, and packet arrival process as Poisson process with average 5 packets/slot, i.e., $\alpha = \alpha_{n_g} = 5$ packets/slot for all n_g unless otherwise stated [14]. The channel gain fading follows Rayleigh distribution, i.e., $h_{n_g,k} = \mathcal{X} d_{n_g}^{-\beta}$, where $\beta = 3$ is path loss exponent, $d_{n_g} \geq 0.35$ is distance between user n_g and BS, normalized to the cell diameter, and $\mathcal{X}$ is exponentially distributed with mean one. The simulations are performed for 1000 on-line time slots and off-line parameters are up-dated after every 25 slots. For off-line phase, we set $\varepsilon_1 = \varepsilon_2 = 10^{-3}$.

Clearly, by increasing R^{rsv}_g and decreasing D_{n_g}, the transmit power should be increased to hold the related constraints. To understand better their effects on the allocated power, performance of the proposed approaches, we consider two

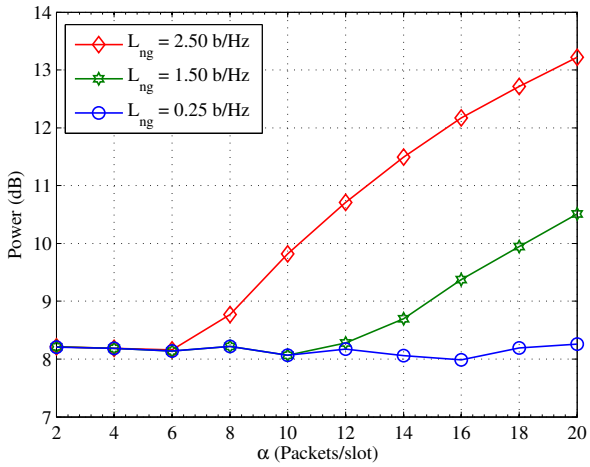


(a) Total transmit power versus D_{n_g} for Case 1

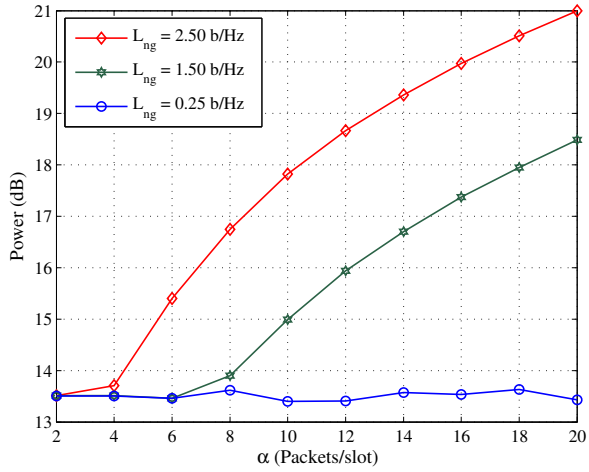


(b) Total transmit power versus D_{n_g} for Case 2

Fig. 1. Total transmit power versus delay D_{n_g} for $G = 2$ slices, $N_1 = N_2 = 2$ users, $K = 64$ sub-carriers with $R_{g1}^{\text{rsv}} = R_{g2}^{\text{rsv}} = 1.0$ bps/Hz and $\alpha = 5$ packets/slot



(a) Total transmit power versus α for Case 1



(b) Total transmit power versus α for Case 2

Fig. 2. Total transmit power versus α for $G = 2$ slices, $N_1 = N_2 = 2$ users, $K = 64$ sub-carriers with $D_{n_g} = 0.5$ and $R_{g1}^{\text{rsv}} = R_{g2}^{\text{rsv}} = 1.0$ bps/Hz

TABLE I
DIFFERENT SETS BASED ON R_g^{RSV} AND L_{n_g}

R_g^{rsv} \ L_{n_g}	2.50 b/Hz	1.50 b/Hz	0.25 b/Hz
2.0 bps/Hz	Set 1	Set 2	Set 3
1.0 bps/Hz	Set 4	Set 5	Set 6
0.5 bps/Hz	Set 7	Set 8	Set 9

different user-location cases: **Case 1** with users close to cell-center, i.e., $d_{n_g} = \{0.35,\ 0.45\}$; and **Case 2** with users close to cell-boundary, i.e., $d_{n_g} = \{0.55,\ 0.65\}$.

The different sets of system parameters, R_g^{rsv} and L_{n_g}, are summarized in Table I. These two parameters determine the lower bound of $\widetilde{\text{C2}}$ and $\widetilde{\text{C3}}$, respectively. By increasing these two parameters, VWN needs more transmit power to satisfy the slice-isolation and/or QoS constraints. Based on our definition, Set 1 has the most stringent constraints while Set 9 has the loosest constraints. Specifically, Sets 1-3 have the strictest $\widetilde{\text{C2}}$ with $R_g^{\text{rsv}} = 2.0$ while Sets 7-9 have relaxed $\widetilde{\text{C2}}$ with $R_g^{\text{rsv}} = 0.5$. Sets 1, 4 and 7 have a larger packet size $L_{n_g} = 2.5$ b/Hz leading to higher Z_{n_g} than Sets 3, 6 and 9 with $L_{n_g} = 0.5$ b/Hz.

Figs. 1(a) and 1(b) depict the transmit power versus D_{n_g} for Case 1 and Case 2, respectively. Both figures show that for sets 3, 6 and 9 with smallest average packet size L_{n_g}, the variations in total transmission power with respect to D_{n_g} is negligible in comparison to the other sets with higher values of L_{n_g}. This is because for the small packet size, D_{n_g} does not have a strict effect on Z_{n_g}, and hence, the total transmit power is robust against variation in D_{n_g}, while for large L_{n_g} (i.e., Sets 1, 4, and 7), D_{n_g} has a strong effect on Z_{n_g}, leading to higher transmit power. Furthermore, it is also observed that

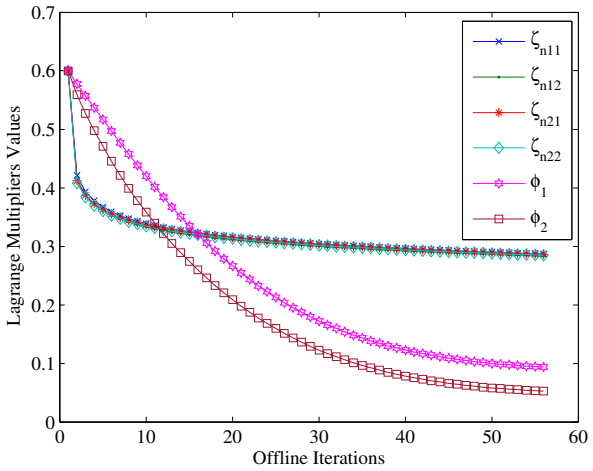


Fig. 3. Convergence of Lagrange Multipliers

with increasing R_g^{rsv}, the transmit power increases to achieve the required minimum rate of slices. The transmit power of Case 2 is larger than Case 1 to compensate the path-loss.

Figs. 2(a) and 2(b) depict the transmit power versus α_{n_g} for Case 1 and Case 2, respectively, where R_{g1}^{rsv} and R_{g2}^{rsv} are set to 1.0 bps/Hz and $D_{ng} = 0.5$. Both figures show that with increasing α_{n_g}, the total transmit power is increased for $L_{n_g} > 0.25$. However, for $L_{n_g} \leq 0.25$, α_{n_g} does not have considerable affect on the transmit power. This is because for large values of packet size ($L_{n_g} > 0.25$), Z_{n_g} increases with increasing α_{n_g} and hence increases the strictness of constraint $\widetilde{\text{C3}}$. As a result, total transmit power requirements of users increases to satisfy the feasibility of $\widetilde{\text{C3}}$. On the other side, for small values of packet size ($L_{n_g} \leq 0.25$), Z_{n_g} does not increase significantly with increasing α_{n_g} and hence $\widetilde{\text{C3}}$ is not affected much. Consequently, for small packet size, increasing α_{n_g}, negligibly affect the total power transmission of users. Therefore, when $\widetilde{\text{C3}}$ is dominant constraint among the other constraints of considered resource allocation problem, increasing α_{n_g} affect the feasibility region of resource allocation problem, and hence, the total transmit power is increased. Whereas, for small packet size, $\widetilde{\text{C3}}$ is no longer dominate as compared to other constraints of the considered problem and the increment of α_{n_g} does not affect the feasibility region.

In summary, Figs. 1-2 show the effect of channel attenuation, i.e., the users with lower channel gains need higher transmit power. In case that the transmit power of users is limited, the admission control policy, e.g., [7], is required to maintain the performance of networks. Otherwise, even with maximum transmit power by users, none of QoS of users and isolation factor between slices can be satisfied.

Finally, we study the convergence of Lagrange variables in off-line phase for 1000 CSI samples generated with prior CDI information. Fig. 3 shows the values of Lagrange multipliers ϕ_g (for $\widetilde{\text{C2}}$) and ζ_{n_g} (for $\widetilde{\text{C3}}$) for all $g \in \mathcal{G}$ and $n_g \in \mathcal{N}_g$ versus off-line iterations with $L_{n_g} = 2.5$ b/Hz. All the Lagrange variables converge to a constant value with-in 60 iterations, indicating the effectiveness of Alg 1.

V. Conclusions

In this work, we propose a delay-aware resource provisioning policy for VWN to minimize total transmit power subject to minimum average rates of each slice and maximum average packet transmission delay of each user. Via effective capacity, we transform all the constraints to the physical-layer dependent constraints, and convexify the formulated cross-layer resource allocation problem with relaxation techniques. Iterative algorithm for joint power and sub-carrier allocation is proposed to solve the proposed delay-aware and power-efficient problem. Via simulation results, the effect of increasing the minimum required rate of slices and decreasing the amount of tolerable delay for each user, the total transmit power of users is incremented.

References

[1] C. Liang, F. Yu, and X. Zhang, "Information-centric network function virtualization over 5G mobile wireless networks," *IEEE Network*, vol. 29, no. 3, pp. 68–74, May 2015.

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

[3] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1333–1346, Oct 2012.

[4] M. Yang, Y. Li, L. Zeng, D. Jin, and L. Su, "Karnaugh-map like online embedding algorithm of wireless virtualization," in *Intl. Symp. on Wireless Personal Multimedia Commun. (WPMC)*, Sept 2012, pp. 594–598.

[5] F. Fu and U. Kozat, "Stochastic game for wireless network virtualization," *IEEE/ACM Trans. Netw.*, vol. 21, no. 1, pp. 84–97, Feb 2013.

[6] M. Yang, Y. Li, J. Liu, D. Jin, J. Yuan, and L. Zeng, "Opportunistic spectrum sharing for wireless virtualization," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, April 2014, pp. 1803–1808.

[7] S. Parsaeefard, V. Jumba, M. Derakhshani, and T. Le-Ngoc, "Joint resource provisioning and admission control in wireless virtualized networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, March 2015, pp. 2020–2025.

[8] V. Jumba, S. Parsaeefard, M. Derakhshani, and T. Le-Ngoc, "Resource provisioning in wireless virtualized networks via massive-MIMO," *IEEE Commun. Lett.*, vol. 4, no. 3, pp. 237–240, June 2015.

[9] D. Hui, V. Lau, and W. H. Lam, "Cross-layer design for OFDMA wireless systems with heterogeneous delay requirements," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 2872–2880, August 2007.

[10] J. Tang and X. Zhang, "Cross-layer resource allocation over wireless relay networks for quality of service provisioning," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 4, pp. 645–656, May 2007.

[11] M. Tao, Y.-C. Liang, and F. Zhang, "Adaptive resource allocation for delay differentiated traffic in multiuser OFDM systems," in *Proc. IEEE Intl. Conf. Commun. (ICC)*, vol. 10, June 2006, pp. 4403–4408.

[12] I. Wong and B. Evans, "Optimal OFDMA resource allocation with linear complexity to maximize ergodic weighted sum capacity," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 3, April 2007, pp. III–601–III–604.

[13] I. Wong and B. Evans, "Optimal downlink OFDMA resource allocation with linear complexity to maximize ergodic rates," *IEEE Trans. Wireless Commun.*, vol. 7, no. 3, pp. 962–971, March 2008.

[14] Y. Cui, V. Lau, R. Wang, H. Huang, and S. Zhang, "A survey on delay-aware resource control for wireless systems-large deviation theory, stochastic lyapunov drift, and distributed stochastic learning," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1677–1701, March 2012.

[15] D. Ng and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," in *Proc. IEEE Intl. Conf. Commun. (ICC)*, June 2012, pp. 5916–5920.

[16] Y. Wang, J. Zhang, and P. Zhang, "Energy-efficient power and subcarrier allocation in multiuser OFDMA networks," in *Proc. IEEE Intl. Conf. Commun. (ICC)*, June 2014, pp. 5492–5496.
[17] D. Ng, E. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, September 2012.
[18] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[19] X. Wang and G. Giannakis, "Resource allocation for wireless multiuser OFDM networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4359–4372, July 2011.
[20] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, July 2006.