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

This item was submitted to Loughborough University's Institutional Repository by the/an author.

Citation: PARASEEFARD, S. ... et al., 2016. Delay-aware and power-efficient resource allocation in virtualized wireless networks. IN: Proceedings of 2016 IEEE Wireless Communications and Networking Conference (WCNC), 3-6 April 2016, DOI: 10.1109/WCNC.2016.7564648

Additional Information:

- © 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Metadata Record: https://dspace.lboro.ac.uk/2134/21334

Version: Accepted for publication

Publisher: © IEEE

Rights: This work is made available according to the conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. Full details of this licence are available at: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the published version.
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 $K = \{1, \cdots, K\}$ 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 $\mathbf{h} = [h_{n_g,k}]_{N_g \times N_g \times K}$ 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

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

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 $R_{n_g,k} = \log_2(1 + p_{n_g,k}h_{n_g,k})$, where $\sigma$ is the noise power in each sub-carrier and users. Consequently, the total achievable rate of user $n_g$ becomes $R_{n_g}(P, W) = \sum_{k \in \mathcal{K}} R_{n_g,k}$, where $P = \{p_{n_g,k}\}_{\forall n_g, k \in \mathcal{K}}$ and $W = \{w_{n_g,k}\}_{\forall n_g, k \in \mathcal{K}}$ 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 $R_{g}^{sv}$ i.e.,

$$C_2: \mathbb{E}_h \left\{ \sum_{n_g \in \mathcal{N}_g} R_{n_g}(P, W) \right\} \geq R_{g}^{sv}, \quad \forall g \in \mathcal{G},$$

where operator $\mathbb{E}_h[\cdot]$ represents the expectation over random vector $h$.

- **User QoS constraints:** by keeping the average traffic delay for each user $n_g$ for all $g \in \mathcal{G}$ below the predefined threshold, $\alpha_{n_g}$ (Eq. 17 in [14]), i.e.,

$$C_3: \alpha_{n_g} \mathbb{E}_h \left\{ Q_{n_g} \right\} \leq D_{n_g}, \quad \forall n_g \in \mathcal{N}_g, \quad \forall g \in \mathcal{G},$$

where $Q_{n_g}$ and $\alpha_{n_g}$ 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

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

subject to $C_1$-$C_3$.

Note that the proposed resource provisioning problem (1) contains discrete and continuous variables such as $W$ and $P$, 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 $P$ and $W$ instead of $Q_{n_g}$ and $\alpha_{n_g}$. In this context, according to Lemma 1 in [9], C3 can be rewritten as

$$\tilde{C}_3: \mathbb{E}_h \left\{ R_{n_g}(P, W) \right\} \geq Z_{n_g}, \quad \forall n_g \in \mathcal{N}_g, \quad \forall g \in \mathcal{G},$$

where

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

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, $C_1$ is changed to

$$\tilde{C}_1: \sum_{g \in \mathcal{G}} \sum_{n_g \in \mathcal{N}_g} w_{n_g,k} \leq 1 \quad \text{and} \quad w_{n_g,k} \in [0, 1], \forall k \in \mathcal{K}. $$

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

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

where $X$ 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, $C_2$ and $C_3$ become

$$\tilde{C}_2: \mathbb{E}_h \left\{ \sum_{n_g \in \mathcal{N}_g} \tilde{R}_{n_g}(X, W) \right\} \geq R_{g}^{sv}, \quad \forall g \in \mathcal{G} \text{ and}$$

$$\tilde{C}_3: \mathbb{E}_h \left\{ \tilde{R}_{n_g}(X, W) \right\} \geq Z_{n_g}, \quad \forall n_g \in \mathcal{N}_g, \quad \forall g \in \mathcal{G},$$

respectively. Consequently (1) can be rewritten as

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

subject to $\tilde{C}_1$-$\tilde{C}_3$.

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

$$L(\phi_g, \zeta_{n_g}, \rho_k, X, W) = -\mathbb{E}_h \left\{ \sum_{g \in \mathcal{G}} \sum_{n_g \in \mathcal{N}_g} \sum_{k \in \mathcal{K}} x_{n_g,k} \right\} \quad \text{(3)}$$

$$+ \sum_{g \in \mathcal{G}} \phi_g \mathbb{E}_h \left\{ \sum_{n_g \in \mathcal{N}_g} \tilde{R}_{n_g} - P_{g}^{sv} \right\}$$

$$+ \sum_{g \in \mathcal{G}} \sum_{n_g \in \mathcal{N}_g} \zeta_{n_g} \mathbb{E}_h \left\{ \tilde{R}_{n_g} - Z_{n_g} \right\}$$

$$+ \sum_{k \in \mathcal{K}} \rho_k \left( 1 - \sum_{g \in \mathcal{G}} \sum_{n_g \in \mathcal{N}_g} w_{n_g,k} \right),$$

where $\rho_k$ for all $k \in \mathcal{K}$, $\phi_g$ for all $g \in \mathcal{G}$ and $\zeta_{n_g}$ for all $n_g \in \mathcal{N}_g$ are the positive Lagrange variables of $C_1$, $C_2$ and $C_3$, respectively. Let $\rho$, $\phi$ and $\zeta$ be the vectors of the corresponding Lagrange variables $\rho_k$, $\phi_g$ and $\zeta_{n_g}$, respectively.
Now, the Dual function related to (3) is

\[ D(\phi, \zeta, \rho) = \max_{X,W} L(\phi, \zeta, \rho, X, W) \]

and consequently, the dual problem is

\[ \min_{\phi, \zeta, \rho} D(\phi, \zeta, \rho) \]

subject to \( C1 - C3 \).

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

\[ p_{n_g,k}^* = \left[ \frac{\phi_g + \zeta_{n_g}}{\ln(2)} - \frac{\sigma}{h_{n_g,k}} \right]_{0}^{\text{max}}, \]

where \( \lceil x \rceil = \max\{\min\{x, a\}, b\} \). From KKT conditions, for optimal sub-carrier allocation \( w_{n_g,k}^* \), we have,

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

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

\[ w_{n_g,k}^* = \begin{cases} 
1, & n_g = \max_{\forall n_g \in N_g, \forall g \in G} \frac{\partial L(\phi_g, \zeta_{n_g}, \rho_k, X, W)}{\partial w_{n_g,k}^*}, \\
0, & n_g \neq n_g^*.
\end{cases} \]

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 \( \phi \) and \( \zeta \) are updated as

\[ \phi_g^*(j) = \phi_g(j - 1) + \delta_{\phi_g} \frac{\partial L}{\partial \phi_g}, \forall g \in G \] and

\[ \zeta_{n_g}^*(j) = \zeta_{n_g}(j - 1) + \delta_{\zeta_{n_g}} \frac{\partial L}{\partial \zeta_{n_g}}, \forall n_g \in N_g, \forall g \in G, \]

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

\[ \phi_g^*(j) = \phi_g(j - 1) + \delta_{\phi_g} \left( \sum_{n_g \in N_g} \tilde{R}_{n_g}(X(j - 1), W(j - 1)) - R_{n_g}^{\text{NW}} \right), \]

\[ \zeta_{n_g}^*(j) = \zeta_{n_g}(j - 1) + \delta_{\zeta_{n_g}} \left( \tilde{R}_{n_g}(X(j - 1), W(j - 1)) - Z_{n_g} \right). \]

The iterative process will be terminated if

\[ \|\phi_g^*(j) - \phi_g(j - 1)\| \leq \varepsilon_1 \text{ and } \|\zeta_{n_g}^*(j) - \zeta_{n_g}(j - 1)\| \leq \varepsilon_2 \]

where \( 0 < \varepsilon_1 < 1 \) and \( 0 < \varepsilon_2 < 1 \).

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 \( j = 0, \phi(j) = 0 \), \( \zeta(j) = 0 \), \( \delta_{\phi_g}(j = 0) \), \( \delta_{\zeta_{n_g}}(j = 0) \), and \( J_{\text{max}} \gg 1 \), as the maximum number of iteration for offline phase.

**Repeat:**

Update \( \phi(j) \) and \( \zeta(j) \) according to (7) and (8).

Until \( (9) \) holds or \( j \geq J_{\text{max}} \).

**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 update \( \phi \) and \( \zeta \) [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 \( R_{g1} = R_{g2} = 1.0 \) 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., \( h_{n_g,k} = \mathcal{X}d_{n_g,k}^\beta \), where \( \beta = 3 \) is path loss exponent, \( d_{n_g,k} \geq 0.35 \) 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 updated after every 25 slots. For off-line phase, we set \( \varepsilon_1 = \varepsilon_2 = 10^{-3} \).

Clearly, by increasing \( R_{g1}^{\text{NW}} \) 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
Fig. 1. Total transmit power versus delay $D_{n_g}$ for Case 1

Fig. 2. Total transmit power versus $\alpha$ for Case 1

<table>
<thead>
<tr>
<th>$R_{gy}$ (bps/Hz)</th>
<th>$L_{n_g}$ (b/Hz)</th>
<th>$2.50$ b/Hz</th>
<th>$1.50$ b/Hz</th>
<th>$0.25$ b/Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2.0$</td>
<td>Set 1</td>
<td>Set 2</td>
<td>Set 3</td>
<td></td>
</tr>
<tr>
<td>$1.0$</td>
<td>Set 4</td>
<td>Set 5</td>
<td>Set 6</td>
<td></td>
</tr>
<tr>
<td>$0.5$</td>
<td>Set 7</td>
<td>Set 8</td>
<td>Set 9</td>
<td></td>
</tr>
</tbody>
</table>

TABLE I
DIFFERENT SETS BASED ON $R_{gy}$ AND $L_{n_g}$

The different sets of system parameters, $R_{gy}$ and $L_{n_g}$, are summarized in Table I. These two parameters determine the lower bound of C2 and 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 C2 with $R_{gy} = 2.0$ while Sets 7-9 have relaxed C2 with $R_{gy} = 0.5$. 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

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, $R_{gy}$ and $L_{n_g}$, are summarized in Table I. These two parameters determine the lower bound of C2 and 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 C2 with $R_{gy} = 2.0$ while Sets 7-9 have relaxed C2 with $R_{gy} = 0.5$. 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
with increasing $R_{gsv}$, 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 $\alpha_{ng}$ for Case 1 and Case 2, respectively, where $R_{gsv}^1$ and $R_{gsv}^2$ are set to 1.0 bps/Hz and $D_{ng} = 0.5$. Both figures show that with increasing $\alpha_{ng}$, the total transmit power is increased for $L_{ng} > 0.25$. However, for $L_{ng} \leq 0.25$, $\alpha_{ng}$ does not have considerable effect on the transmit power. This is because for large values of packet size ($L_{ng} > 0.25$), $Z_{ng}$ increases with increasing $\alpha_{ng}$ and hence increases the strictness of constraint $C_3$. As a result, total transmit power requirements of users increases to satisfy the feasibility of $C_3$. On the other side, for small values of packet size ($L_{ng} \leq 0.25$), $Z_{ng}$ does not increase significantly with increasing $\alpha_{ng}$ and hence $C_3$ is not affected much. Consequently, for small packet size, increasing $\alpha_{ng}$ negligibly affect the total power transmission of users. Therefore, when $C_3$ is dominant constraint among the other constraints of considered resource allocation problem, increasing $\alpha_{ng}$ affect the feasibility region of resource allocation problem, and hence, the total transmit power is increased. Whereas, for small packet size, $C_3$ is no longer dominate as compared to other constraints of the considered problem and the increment of $\alpha_{ng}$ 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 $\phi_g$ (for C2) and $\xi_{ng}$ (for C3) for all $g \in G$ and $ng \in N_g$ versus off-line iterations with $L_{ng} = 2.5$ 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


