Utility Analysis of Radio Access Network Slicing

Guorong Zhou, Liqiang Zhao, *Member, IEEE*, Kai Liang, *Member, IEEE*, Gan Zheng, *Senior Member, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—The utility optimization problem of radio access network-slicing aided mobile systems is formulated considering both throughput and delay demands from a mobile network operator's perspective, whilst relying on stochastic geometry and Lyapunov optimization. Then a joint virtual resource optimization algorithm is proposed to dynamically allocate both virtual spectral and the power resources. Our numerical results support both the high-throughput slice and the low-delay slice and quantify the associated throughput vs. delay trade-off. Moreover, we compare the proposed algorithm with the benchmark one, which ensures better utility.**

***Index Terms*—radio access network slicing, utility, stochastic geometry, Lyapunov optimization.**

I. Introduction

As a salient 5G networking technique, network slicing [1], [2] is proposed for supporting diverse customized services with various resource constraints. The Radio Access Network (RAN) slicing [3], [4] constitutes an important part of the end-to-end network slicing, which is capable of dynamically allocating the RAN resources, such as the virtual base stations (vBSs), as well as the spectral and power resources.

There is rich literature on the resource management and performance analysis of RAN slicing [5]–[9]. Specifically, Sallent *et al.* [5] analyze the radio resource management functionalities of the RAN slicing in a multi-cell network, which can be used for splitting the radio resources among the RAN slices. Moreover, there are numerous performance metrics for characterizing a virtual slicing aided network. For example, Zhang *et al.* [6] consider the flow-rate over the link as the optimization objective function of network slicing. Shi *et al.* [7] carry out the tradeoff analysis between energy efficiency and delay in wireless network virtualization, whilst guaranteeing the users' quality of service. However, with the diverse requirements of the slices, it is a challenge to comprehensively characterize several slices using only a single performance metric.

In fact, RAN slices may also be considered as a set of virtual sub-networks based on the same physical network, and ideally a unified criterion is needed to describe the performance of the entire network. Therefore, Caballero *et al.* [8] define the concept of utility gains for quantifying the users' rate improvement gleaned from dynamic resource sharing across RAN slices. Similarly, Ye *et al.* [9] study the optimal network utility dependence on the different traffic statistics in a dynamic radio resource slicing aided framework. However, there is a paucity of literature on the utility analysis considering both throughput and delay demands in RAN slices, which is hence the focus of our work.

Firstly, we will study the optimal utility of both the high-throughput and the low-delay slices from the perspective of mobile network operator (MNO) by invoking stochastic geometry theory. Secondly, by using Lyapunov optimization, a joint virtual resource optimization algorithm is proposed for maximizing the utility, which dynamically allocates both virtual spectral resources and the vBSs' power. Finally, numerical results will demonstrate that a pair of customized RAN slices can be supported, which ensures better utility than the benchmark algorithm. Simultaneously, there is the tradeoff between delay and throughput in both the slices once the QoS requirement is satisfied.

II. System Model

A. High-throughput Slice and Low-delay Slice

Without loss of generality, we consider the downlink of a RAN, in which N physical base stations (BSs) and multiple user equipment (UE) are randomly distributed in the Euclidean plane having an area of S_{area}. We rely on the homogeneous Possion Point Process (PPP) ϕ_B having a density of λ_B and ϕ_k to model the BSs and UEs deployment respectively [10], where $N = \lambda_B S_{area}$. Furthermore, a buffer is attached to each physical BS for the data queues. The bandwidth of the system is B and the power of each physical BS is P_B. We define the time slots as a time interval $[t, t+1)$, where $t \in \{0, 1, 2, ...\}$ and the duration of every slot is τ.

In this system, the MNO needs to provide a pair of slices for the users' high-throughput and low-delay services. Firstly, a physical BS may be mapped into one or two vBSs, each of which is associated with a specific service by virtualization. And the communication resources in the system are abstracted into virtual resources to realize sharing. Next, the MNO determines the vBS deployment as well as the vBS's power and virtual spectrum allocation for each service at slot t for creating a high-throughput slice and a low-delay slice, defined as Slice 1 and Slice 2, respectively.

Specifically, we define the activation probability $p_s(t)$[1] to indicate a particular vBS's deployment in slice s $(s \in \{1, 2\})$ at slot t, the vBS's power allocation factor $\eta_s(t)$ as the vBS's power ratio of slice s at slot t and the virtual spectrum allocation factor $\rho_s(t)$ as the ratio of spectral resources allocated

Guorong Zhou, Liqiang Zhao and Kai Liang are with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China (e-mail: guor_zhou@163.com; lqzhao@mail.xidian.edu.cn; kliang@xidian.edu.cn). Gan Zheng is with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough LE11 3TU, U.K. (e-mail: g.zheng@lboro.ac.uk). Lajos Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

[1] Note that s $(s \in \{1, 2\})$ in all symbols represents slice 1 and slice 2.

to slice s at slot t. Since the two slices are independent of each other, the location of the vBSs and of the UEs in the RAN slice s at slot t can be represented by the dynamic homogeneous thinning PPP ϕ_s^{VB} associated with the density of $\lambda_s^{VB}(t) = p_s(t)\lambda_B$ and ϕ_s^k with the density of $\lambda_s^k(t)$, where $p_s(t) \in (0,1]$. Furthermore, we can express the power of each vBS in slice s at slot t, which is given as $P_s^{VB}(t) = \frac{\eta_s(t)NP_B}{\lambda_s^{VB}(t)S_{area}} = \frac{\eta_s(t)}{p_s(t)}P_B$. Additionally, we have $\rho_s(t) = \frac{\lambda_s^k(t)b_s(t)}{\lambda_s^{VB}(t)B}$, where $b_s(t)$ indicates the spectra allocated to each user in slice s at slot t by the MNO, and $\lambda_s^k(t)/\lambda_s^{VB}(t)$ represents the number of users associated with a vBS in slice s at slot t. Hence the bandwidth allocated to slice s at slot t is $\rho_s(t)B$. Let us assume that there is no intra-cell interference, and that the bandwidths assigned to the two slices do not overlap with each other in every slot, hence there is no interference between the slices, which is an explicit benefit of their isolation.

B. Utility Definition of RAN Slices

Without loss of generality, each UE in slice s is associated with the closest vBS at slot t and $r_s^{k,min}(t)$ represents the distance between them. Hence, for a typical user k in slice s at slot t, the signal to interference plus noise ratio (SINR) can be expressed as:

$$SINR_s^k(t) = \frac{P_s^{VB}(t)h_s^{k,VB_i}(t)\left(r_s^{k,min}(t)\right)^{-\alpha}}{I_s^k(t)+\sigma^2}, \tag{1}$$

where $h_s^{k,VB_i}(t)$ is the channel gain between user k and its nearest vBS VB_i in slice s at slot t, which follows an exponential distribution with unit mean, α denotes the path loss exponent, while σ^2 is the noise power, and $I_s^k(t)$ is the interference arriving from other vBSs in slice s at slot t, which is expressed as:

$$I_s^k(t) = \sum\nolimits_{j\in\phi_s^{VB}, j\neq i} P_s^{VB}(t)h_s^{k,VB_j}(t)\left(r_s^{k,VB_j}(t)\right)^{-\alpha}. \tag{2}$$

To indicate the vBS's dynamic deployment density, the activation probability $p_s(t)$ of the vBSs in slice s at slot t is evaluated according to the users' SINR status in the current slot, defined as the probability that the users in slice s achieve the current target SINR T_s. As a special case, when $\alpha=4$, the activation probability can be derived as:

$$\begin{aligned}
&p_s(t) = p_s\left[T_s, \eta_s(t), \rho_s(t)\right]\\
&= E_r\left\{P\left[SINR_s^k(t) > T_s \middle| r_s^{k,min}(t)\right]\right\}\\
&= \int_0^\infty P\left[h_s^{k,VB_i} > \frac{T_s\left(r_s^{k,min}(t)\right)^\alpha\left(I_s^k(t)+\sigma^2\right)}{P_s^{VB}(t)} \middle| r_s^{k,min}(t)\right]\\
&\cdot e^{-\pi\lambda_s^{VB}(t)\left(r_s^{k,min}(t)\right)^2} 2\pi\lambda_s^{VB}(t)r_s^{k,min}(t)dr_s^{k,min}(t)\\
&\overset{\alpha=4}{=} \sqrt{\frac{\pi}{e_s(t)}}\exp\left(\frac{{d_s}^2(t)}{4e_s(t)}\right)Q\left(\frac{d_s(t)}{\sqrt{2e_s(t)}}\right),
\end{aligned} \tag{3}$$

where $d_s(t) = 1+\rho_s(t)\sqrt{T_s}\left(\frac{\pi}{2} - \arctan\left(\frac{1}{\sqrt{T_s}}\right)\right)$, $e_s(t) = \frac{\sigma^2 T_s}{\pi^2 p_s(t)\eta_s(t)P_B\lambda_B^2}$ and $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left(-y^2/2\right)dy$ represents the standard Gaussian tail probability. As we can see, $p_s(t)$ is inversely proportional to $\eta_s(t)$ and $\rho_s(t)$, respectively.

According to the classic Shannon formula and stochastic geometry theory [10], the data rate $R_s^k(t)$ that the RAN slice s can provide for user k at slot t when $\alpha=4$ is given as:

$$\begin{aligned}
&R_s^k(t) = b_s(t)\cdot E\left[\log_2\left(1+SINR_s^k(t)\right)\right]\\
&\overset{\alpha=4}{=} \frac{b_s(t)}{\ln 2}\int_0^\infty \sqrt{\frac{\pi}{c_s(x,t)}}\exp\left(\frac{a_s^2(x,t)}{4c_s(x,t)}\right)Q\left(\frac{a_s(x,t)}{\sqrt{2c_s(x,t)}}\right)dx,
\end{aligned} \tag{4}$$

where $a_s(x,t) = 1+\rho_s(t)\sqrt{e^x-1}\left(\frac{\pi}{2} - \arctan\left(\frac{1}{\sqrt{e^x-1}}\right)\right)$ and $c_s(x,t) = \frac{\sigma^2(e^x-1)}{P_s^{VB}(t)(\pi\lambda_s^{VB}(t))^2}$. Consequently, the downlink throughput in RAN slice s at slot t may be expressed as:

$$\begin{aligned}
R_s^{sum}(t) &= R_s^{sum}\left[\eta_s(t), \rho_s(t), \lambda_s^k(t)\right]\\
&= \lambda_s^k(t)S_{area}\left[R_s^k(\eta_s(t), \rho_s(t))\right].
\end{aligned} \tag{5}$$

The time-averaged expectation of the throughput can be defined as

$$\overline{R}_s^{sum}\left[\eta_s(t), \rho_s(t)\right] = \lim_{t\to\infty}\frac{1}{t}\sum\nolimits_{\tau=0}^{t-1}\mathrm{E}\left\{R_s^{sum}(\tau)\right\}. \tag{6}$$

At the same time, we can express the power consumed at slot t by the physical network as:

$$\begin{aligned}
&P_{sum}(t) = P_{sum}\left[\eta_1(t), \eta_2(t)\right]\\
&= P_1^{VB}(t)\lambda_1^{VB}(t)S_{area} + P_2^{VB}(t)\lambda_2^{VB}(t)S_{area}\\
&= \left[\eta_1(t)+\eta_2(t)\right]NP_B.
\end{aligned} \tag{7}$$

The bandwidth used at slot t is:

$$B_{sum}(t) = B_{sum}\left[\rho_1(t), \rho_2(t)\right] = \left[\rho_1(t)+\rho_2(t)\right]B. \tag{8}$$

Next we analyze the delay of RAN slice s. Let the vectors $A_s(t) = \{A_s^1(t), A_s^2(t), ..., A_s^k(t), ...\}$ and $Q_s(t) = \{Q_s^1(t), Q_s^2(t), ..., Q_s^k(t), ...\}$ represent the processes of random data arrivals and current data queue backlogs in RAN slice s at slot t, respectively, where $A_s(t)$ is independent and identically distributed (i.i.d.) over time, as governed by the arrival rate γ_s. Naturally, the data arrival rate of slice 1 is higher than that of slice 2, i.e. $\gamma_1 > \gamma_2$. We model the queuing process of the data requested by user k of slice s at slot t as

$$Q_s^k(t+1) = \max[Q_s^k(t) - R_s^k(t)\tau, 0] + \mathrm{A}_s^k(\mathrm{t}), \forall k, s. \tag{9}$$

A network is stable, when all these discrete queues $Q_s^k(t)$ are mean rate stable, i.e. they satisfy the following condition [12]:

$$\lim_{t\to\infty}\frac{E\{|Q_s^k(t)|\}}{t} = 0. \tag{10}$$

When the network is stable, based on Little's Theorem [11], we can get the average delay of the two slices according to the average data queue length:

$$\overline{D}_s^k = \overline{Q}_s^k/\gamma_s, \tag{11}$$

where $\overline{Q}_s^k = \lim\limits_{t\to\infty}\frac{1}{t}\sum\limits_{\tau=0}^{t-1}\mathrm{E}\left\{Q_s^k(\tau)\right\}$ is the time-averaged data queue length of user k in slice s. Furthermore, $\overline{D}_s^k = \lim\limits_{t\to\infty}\frac{1}{t}\sum\limits_{\tau=0}^{t-1}\mathrm{E}\left\{D_s^k(\tau)\right\}$ is the time-averaged delay of user k in slice s and $D_s^k(t)$ is the queuing delay of user k at slot t in slice s.

From the MNO's perspective, utility is defined as the difference between the income represented by the gain gleaned and the cost of the entire network. More specifically, for the throughput-guaranteed slice 1, we define the utility as the difference between the income defined by the throughput gain and the cost quantified in terms of the power and spectral resources in slice 1, formulated as:

$$U_1\left(t\right) = m_1 R_1^{sum}\left(t\right) - \left[\beta \eta_1\left(t\right) N P_B + \delta \rho_1\left(t\right) B\right], \qquad (12)$$

where m_1 is the unit price of throughput gain charged by the MNO, while β and δ are the power and spectral resource unit prices, respectively. For the delay-guaranteed slice 2, we define the utility as the income defined by the delay gain minus the cost in terms of the power and spectral resources in slice 2, namely:

$$U_2(t) = \sum\nolimits_{k=1}^{K_2(t)} \left[\psi - m_2 Q_2^k(t)\right] - \left[\beta \eta_2(t)\, N P_B + \delta \rho_2(t)\, B\right], \quad (13)$$

where m_2 is the unit price of delay gain charged by the MNO, while ψ is the initial maximum benefit of slice 2 [13]. Furthermore, $K_2(t) = \lambda_2^k(t) S_{area}$ denotes the total number of users in slice 2 at slot t. Therefore, we have the system-wide utility for the whole network, which is the sum of the utility for each slice:

$$\begin{aligned} & U\left(t\right) = U_1\left(t\right) + U_2\left(t\right) \\ & = \left\{ m_1 R_1^{sum}\left(t\right) + \sum\nolimits_{k=1}^{K_2(t)} \left[\psi - m_2 Q_2^k\left(t\right)\right] \right\} \\ & - \left[\beta P_{sum}\left(t\right) + \delta B_{sum}\left(t\right)\right]. \end{aligned} \qquad (14)$$

Similarly, the time-averaged expectation of the utility is given by

$$\overline{U}\left[\eta_1\left(t\right), \eta_2\left(t\right), \rho_1\left(t\right), \rho_2\left(t\right)\right] = \lim_{t \to \infty} \frac{1}{t} \sum\nolimits_{\tau=0}^{t-1} \mathrm{E}\left\{U\left(\tau\right)\right\}. \quad (15)$$

III. Problem Formulation And Solution

In this section, we develop the optimal utility under specific throughput and delay constraints. Mathematically, the problem can be formulated as

$$\begin{aligned} \max \quad & \overline{U}\left(\eta_1\left(t\right), \eta_2\left(t\right), \rho_1\left(t\right), \rho_2\left(t\right)\right) \\ \text{s.t.} \quad & C1: \overline{R}_s^{sum} \geq R_s^{av}, \forall s, \\ & C2: \overline{D}_s^k \leq \omega_s, \forall k, s, \\ & C3: \eta_1(t) + \eta_2(t) \leq 1, \eta_s(t) \in \left(0, 1\right), \forall t, s, \\ & C4: \rho_1(t) + \rho_2(t) \leq 1, \rho_s(t) \in \left(0, 1\right), \forall t, s, \end{aligned} \qquad (16)$$

where R_s^{av} and ω_s denote the minimal average throughput and the maximal tolerable delay requested by the users in slice s, respectively. To elaborate, $C1$ is used to satisfy the average throughput R_s^{av} requested by users, while $C2$ guarantees the stability of the queues and limits the average delay in RAN slices. Furthermore, $C3$ and $C4$ represent the non-negativity and the value range of the vBS's power allocation factor and the virtual spectrum allocation factor, respectively.

Based on the general Lyapunov theory [12], we transform the average rate requirement $C1$ in (16) into a virtual rate queue stability problem $\tilde{C}1$ in (19). The virtual rate queue is defined as $G_s^k(t)$, where $G_s^k(0) = 0$ and

$$G_s^k(t+1) = \max[G_s^k(t) + R_s^{av} - R_s^k(\mathrm{t}), 0], \forall k, t, s. \qquad (17)$$

By jointly taking into account the constraint $C2$ and the data queuing process (9), when the delay exceeds ω_s, the received information becomes stale. Hence the network temporarily refuses to process new arrivals. Then the constraint $C2$ in (16) can be equivalently written as $\tilde{C}2$:

$$Q_s^k(t+1) = \begin{cases} \max[Q_s^k(t) - R_s^k(t)\tau, 0] + A_s^k(t), & \text{if } \overline{D}_s^k \leq \omega_s, \\ \max[Q_s^k(t) - R_s^k(t)\tau, 0], & \text{otherwise.} \end{cases} \qquad (18)$$

Then, (16) can be equivalently reformulated as:

$$\begin{aligned} \max \quad & \overline{U}\left[\rho_1\left(t\right), \rho_2\left(t\right), \eta_1\left(t\right), \eta_2\left(t\right)\right] \\ \text{s.t.} \quad & \tilde{C}1: G_s^k\left(t\right) \text{ is mean rate stable}, \forall k, s, \\ & \tilde{C}2, C3, C4. \end{aligned} \qquad (19)$$

To tackle $\tilde{C}1$ and $\tilde{C}2$ in (18), a combined vector, $\Theta(t) = [Q_s(t), G_s(t)]$, is defined for representing the queuing states of all queues, where $Q_s(t)$ and $G_s(t)$ are virtual queue sets.

Then, we can get the conditional Lyapunov drift $\triangle[\Theta(t)]$:

$$\Delta[\Theta(t)] = E\{L[\Theta(t+1)] - L[\Theta(t)] \,|\, \Theta(t)\}, \qquad (20)$$

where $L[\Theta(t)]$ is the Lyapunov function given by

$$L[\Theta(t)] = \frac{1}{2} \sum\nolimits_{s=1}^{2} \sum\nolimits_{k=1}^{K_s(t)} \left[\left(Q_s^k(t)\right)^2 + \left(G_s^k(t)\right)^2\right]. \qquad (21)$$

The drift-plus-penalty expression here can be obtained as

$$F(t) = \Delta[\Theta(t)] - V \cdot E[U(t)], \qquad (22)$$

where $V \geq 0$ is the weight factor between the utility and delay.

We can then express the upper bound of $F(t)$ by invoking Lyapunov's optimization theory [12] as:

$$\begin{aligned} & F(t) \leq B + \sum\nolimits_{s=1}^{2} \sum\nolimits_{k=1}^{K_s(t)} \left\{\xi_s Q_s^k(t) E\left[A_s^k(t) | \Theta(t)\right]\right\} \\ & + \sum\nolimits_{s=1}^{2} \sum\nolimits_{k=1}^{K_s(t)} \left\{G_s^k(t) E\left[R_s^{av} - R_s^k\left(t\right) | \Theta(t)\right]\right\} - \\ & \sum\nolimits_{s=1}^{2} \sum\nolimits_{k=1}^{K_s(t)} \left\{Q_s^k(t) E\left[R_s^k\left(t\right) \tau | \Theta(t)\right]\right\} - V \cdot E\left[U(t) | \Theta(t)\right], \end{aligned} \qquad (23)$$

where ξ_s is a function of $Q_s^k(t)$. If $Q_s^k(t) \leq \omega_s$, $\xi_s = 1$; otherwise, $\xi_s = 0$, which satisfies $\tilde{C}2$ in (19). Furthermore, $B > 0$ is a constant.

Specifically, the optimal solution of the above problem in (19) can be obtained by minimizing the upper bound of $F(t)$ slot by slot with the aid of stochastic optimization theory [12], i.e., by solving the problem below:

$$\begin{aligned} & \min \sum\nolimits_{s=1}^{2} \sum\nolimits_{k=1}^{K_s(t)} \left[\xi_s Q_s^k(t) A_s^k(t) + G_s^k(t) \left(R_s^{av} - R_s^k\left(t\right)\right)\right] \\ & - \sum\nolimits_{s=1}^{2} \sum\nolimits_{k=1}^{K_s(t)} \left[Q_s^k(t) R_s^k\left(t\right) \tau\right] - V\left[U(t)\right] \\ & \text{s.t.} \quad \tilde{C}3: \eta_1(t) + \eta_2(t) = 1, \eta_s(t) \in \left(0, 1\right), \forall t, s, \\ & \qquad \tilde{C}4: \rho_1(t) + \rho_2(t) = 1, \rho_s(t) \in \left(0, 1\right), \forall t, s. \end{aligned} \qquad (24)$$

At this point, we have transformed the challenging original problem (16) into the problem (24) that is easier to solve. However, to ensure that the optimal solution is obtained, the constraint $C3$ and $C4$ must be converted into $\tilde{C}3$ and $\tilde{C}4$, respectively. Therefore, we can propose an efficient joint virtual resource optimization algorithm based on the classic drift-plus-penalty algorithm of [12] as summarized in Algorithm 1. Here we are assuming that the power- and spectral-resources in the network are sufficient to satisfy the users' average data rate and delay requirements. In particular, the solution of (24) can be found using the DIRECT algorithm of [14] at

Algorithm 1 Joint virtual resource optimization algorithm to solve (16).

```
Initialization: $Q_s^k(0)$=0, $G_s^k(0)$=0, and $U(0)$=0, $\forall k, s$.
Repeat:
Update vBS's deployment density $\lambda_s^{VB}(t)$ according to (3).
Update vBS's power allocation factor $\eta_s(t)$ and the virtual spectrum allocation factor $\rho_s(t)$ according to (24).
Let $t = t + 1$.
Update $Q_s^k(t)$, $G_s^k(t)$ and $U(t)$ according to (18), (17) and (14), respectively.
Stop when $t = T$, where $T$ is the total number of time slots.
```

a complexity order of $O(TW^2)$, where $W = \max[\eta_s(t), \rho_s(t)]$ is the maximum of the sampling number $\eta_s(t)$ and $\rho_s(t)$.

Furthermore, the theoretical bounds of the utility can be derived.

Theorem 1: The utility in (16) is bounded by

$$U^{opt} \ge \overline{U} = \overline{U}\left[\eta_1(t), \eta_2(t), \rho_1(t), \rho_2(t)\right] \ge U^{opt} - \frac{B}{V}, \quad (25)$$

where U^{opt} is the theoretical optimal value of $\overline{U}$.

Proof: Assume that $R_s^{k*}(t)$ and $U^*(t)$ result from a feasible virtual resource allocation policy $\{p_s(t), \eta_s(t), \rho_s(t)\}$ of problem (16) at slot t. The following conditions can be obtained for any $\delta > 0$ and $\varepsilon > 0$ according to stochastic optimization theory [12]:

$$E[A_s^k(t) - R_s^{k*}(t)\tau|\Theta(t)] = E[A_s^k(t) - R_s^{k*}(t)\tau] \le -\varepsilon, \quad (26)$$

$$E[U^*(t)|\Theta(t)] = E[U^*(t)] \ge U^{opt} + \delta, \quad (27)$$

$$E[R_s^{k*}(t)|\Theta(t)] = E[R_s^{k*}(t)] \ge \gamma_s + \varepsilon, \quad (28)$$

$$E[R_s^{av} - R_s^{k*}(t)|\Theta(t)] = E[R_s^{av} - R_s^{k*}(t)] \le \delta. \quad (29)$$

Upon substituting (26)-(29) into the right-hand side of (23) and letting $\delta \to 0$, we arrive at the following inequality:

$$\begin{aligned}\Delta[\Theta(t)] - VE[U(t)|\Theta(t)] \le & \sum_{s=1}^{2} \sum_{u=1}^{U(t)} (\xi_s - 1)\, Q_s^k(t)\gamma_s \\ & -\varepsilon \sum_{s=1}^{2} \sum_{k=1}^{K_s(t)} Q_s^k(t) - VU^{opt} + B.\end{aligned} \quad (30)$$

Taking the expectation of both sides of (30), and using the telescope sum over $t \in \{0, 1, ..., H-1\}$ for the result as well as exploiting the fact that $Q_s^k(t) \ge 0$ and $\xi_s - 1 \le 0$, we have

$$E\{L[\Theta(H)]\} - E\{L[\Theta(0)]\} - V\sum_{t=0}^{H-1} E[U(t)] \le HB - HVU^{opt}, \quad (31)$$

where H represents the number of time slots.

Dividing both sides of the above inequality by VH yields the following:

$$\frac{-E\{L[\Theta(0)]\}}{VH} - \frac{1}{H}\sum_{t=0}^{H-1} E[U(t)] \le \frac{B}{V} - U^{opt}. \quad (32)$$

Upon letting $H \to \infty$ and considering the equation $E\{L[\Theta(0)]\} \ge 0$, we arrive at the following conclusion

$$\overline{U} = \frac{1}{H}\sum_{t=0}^{H-1} E[U(t)] \ge U^{opt} - \frac{B}{V}. \quad (33)$$

Thus, (25) is proved. ■

IV. Numerical Results And Discussions

In this section, we provide numerical results for validating the theoretical analysis. We assume that the transmission power of a physical BS is P_B=$40W$ and the bandwidth available in the system is B=$18MHz$. In addition, we set λ_u=$0.00025(/m^2)$, λ_B=$0.00001(/m^2)$, S_{area}=$2km \times 2km$ and $\tau = 1ms/slot$. For ease of analysis, we set the unit prices are m_1=$0.001(s/bits)$, m_2=$1(/bits)$, β=$50(/W)$, δ=$0.005(/Hz)$ and the initial maximum benefit is ψ=800. Furthermore, we set the thresholds to R_1^{av}=$900Mbps$, R_2^{av}=$400Mbps$, ω_1=$10ms$ and ω_2=$2.5ms$. Moreover, we compare our proposed algorithm to the benchmark algorithm of the even resources allocation between two slices [15], which is defined as the "even allocation based algorithm".

Fig. 1 shows several utility curves versus the data arrival rate γ_1 and γ_2. The system-wide utility and the utility of slice 1 are seen to increase near-linearly, but then they tend to gracefully saturate upon further increasing γ_1 beyond 800 Mbps. This indicates that as expected, the MNO could get a higher utility at higher traffic loads. However, if the traffic load is excessive, its utility remains unchanged because of the associated resource limitation. By contrast, as γ_2 increases, the system-wide utility and utility of slice 2 decrease near-linearly. This is because the increase of traffic loads leads to the escalation of delay in slice 2, which has a detrimental impact on both the utilities. Observe that the system-wide utility obtained by the even allocation based algorithm is much lower than that obtained by our proposed algorithm, so our algorithm guarantees that the MNO gets a higher utility. Finally, because slices 1 and 2 are isolated from each other, the traffic load of one slice has little effect on the utility of another slice.

Fig. 2 and Fig. 3 depict the throughput and delay of slice 1 and 2 versus the data arrival rate γ_1 and γ_2, respectively. As expected, upon increasing γ_1, the throughput of slice 1 gradually increases first and then saturates when the users' arrival rate becomes excessive, which is a consequence of the inherent resource limitation. At the same time, the delay of slice 1 increases near-linearly with γ_1 under the limit of ω_1, which indicates that our proposed algorithm gives priority to a high throughput for slice 1 even at a high traffic load. The trends of slice 2 are similar to those of slice 1. However, slice 2 avoids exceeding the max tolerable delay 2.5 ms, which is much lower than the delay of slice 1. The throughput of slice 2 is also much lower than that of slice 1. It is worth noting that since the two slices are isolated from each other, the change of one slice's traffic load does not affect the QoS of the other slice.

Fig. 4 shows the throughput versus average delay of slices 1 and 2. Observe in Fig. 4 that increasing the throughput is always at the expense of increasing the delay of slice 1. Similarly, the minimization of delay is always accompanied by a throughput-reduction in slice 2. There is an inevitable tradeoff between the delay and throughput in both the slices, once the QoS requirement is satisfied.

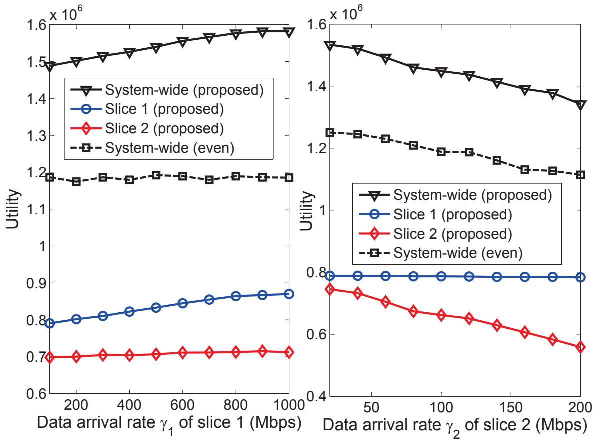

Figure 1. Utilities versus data arrival rate γ_1 and γ_2.

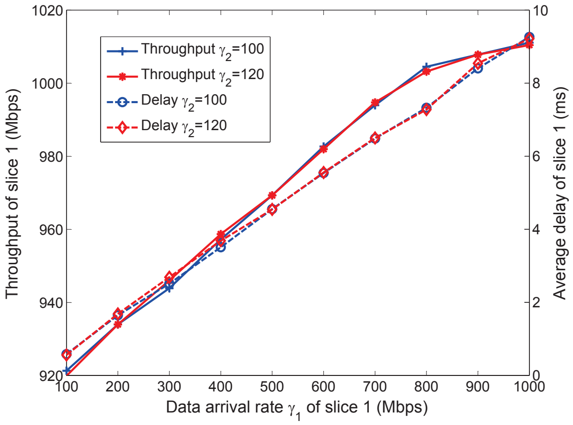

Figure 2. Throughput and average delay of slice 1 versus γ_1.

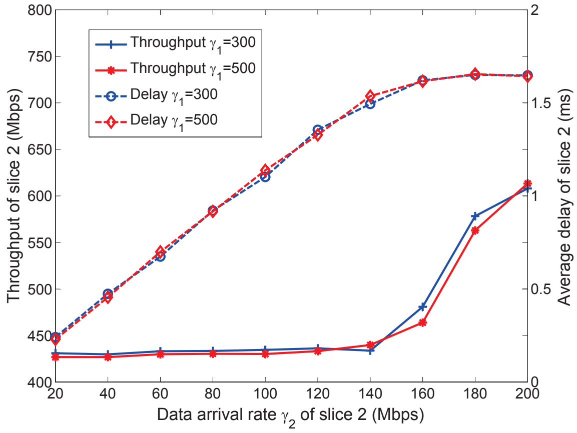

Figure 3. Throughput and average delay of slice 2 versus γ_2.

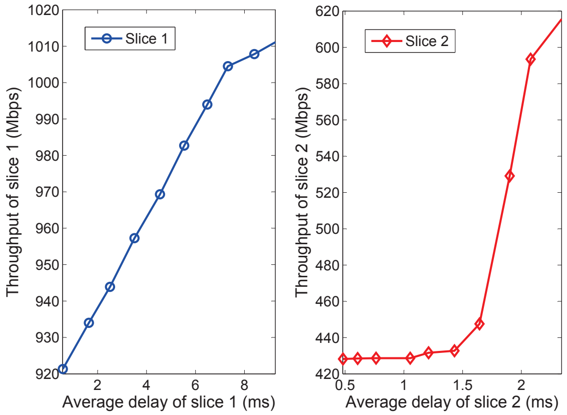

Figure 4. Throughput versus average delay of slices 1 and 2.

V. Conclusions

In this paper, the optimal utility considering both throughput and delay demands of RAN slicing has been analyzed. Firstly, the problem was formulated relying on stochastic geometry theory to maximize the utility. Then, by using Lyapunov optimization, we carried out joint virtual resource optimization by integrating the virtual spectrum and power allocation. Finally, the numerical results verified that both the high-throughput slice and the low-delay slice could be supported and the associated throughput vs. delay trade-off could be struck. Also, our proposed algorithm ensured better utility than the even allocation based algorithm.

References

[1] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini and H. Flinck, “Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429-2453, Third Quarter 2018.

[2] T. Xu, M. Zhang, H. Hu and H. Chen, “Sliced Spectrum Sensing-A Channel Condition Aware Sensing Technique for Cognitive Radio Networks,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10815-10829, Nov. 2018.

[3] K. Koutlia, A. Umbert, S. Garcia and F. Casadevall, “RAN slicing for multi-tenancy support in a WLAN scenario,” 2017 IEEE NetSoft, Bologna, 2017, pp. 1-2.

[4] P. L. Vo, M. N. H. Nguyen, T. A. Le and N. H. Tran, “Slicing the Edge: Resource Allocation for RAN Network Slicing,” *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 970-973, Dec. 2018.

[5] O. Sallent, J. Perez-Romero, R. Ferrus and R. Agusti, “On Radio Access Network Slicing from a Radio Resource Management Perspective,” *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 166-174, October 2017.

[6] N. Zhang, Y. Liu, H. Farmanbar, T. Chang, M. Hong and Z. Luo, “Network Slicing for Service-Oriented Networks Under Resource Constraints,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2512-2521, Nov. 2017.

[7] Q. Shi, L. Zhao, Y. Zhang, G. Zheng, F. R. Yu and H. Chen, “Energy-Efficiency Versus Delay Tradeoff in Wireless Networks Virtualization,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 837-841, Jan. 2018.

[8] P. Caballero, A. Banchs, G. de Veciana and X. Costa-Prez, “Multi-Tenant Radio Access Network Slicing: Statistical Multiplexing of Spatial Loads,” *IEEE/ACM Trans. Networking*, vol. 25, no. 5, pp. 3044-3058, Oct. 2017.

[9] Q. Ye, W. Zhuang, S. Zhang, A. Jin, X. Shen and X. Li, “Dynamic Radio Resource Slicing for a Two-Tier Heterogeneous Wireless Network,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896-9910, Oct. 2018.

[10] J. G. Andrews, F. Baccelli and R. K. Ganti, “A Tractable Approach to Coverage and Rate in Cellular Networks,” *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122-3134, November 2011.

[11] D. P. Bertsekas and R. G. Gallager, *Data Networks (2nd edition)*. Prentice Hall, 1992.

[12] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan and Claypool, 2010.

[13] M. Xiao, J. Wu, C. Liu and L. Huang, “TOUR: Time-sensitive Opportunistic Utility-based Routing in delay tolerant networks,” 2013 Proceedings IEEE INFOCOM, Turin, 2013, pp. 2085-2091.

[14] C.D. Perttunen D.R. Jones and B.E. Stuckman, “Lipschitzian optimization without the lipschitz constant,” *Journal of Optimization Theory and Application*, 79(1):157-181, October 1993.

[15] Y. L. Lee, J. Loo, T. C. Chuah and L. Wang, “Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2146-2161, April 2018.