Rate-delay analysis of radio access network slices

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

Citation: ZHOU, G. ... et al, 2019. Rate-delay analysis of radio access network slices. Presented at the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9-13 December 2018.

Additional Information:

- © 2018 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/2134/37584

Version: Accepted for publication

Publisher: © IEEE

Please cite the published version.
Abstract—Based on wireless network virtualization, radio access network (RAN) slicing is developed to provide services for the different users’ requirements. Moreover, the users’ sum data rate and delay are two significant metrics to guarantee quality of services. In this paper, we first establish an optimization problem to maximize the downlink sum rate while guaranteeing users’ delay for RAN slices, where the base stations and user equipments are randomly distributed. Then we analyze the performance tradeoff between the sum rate maximization and delay tolerance. With the aid of Lyapunov optimization, the upper bounds of the achievable rate and delay are derived, through which the existence of tradeoff in performance is obvious and verified by numerical results.

Index Terms—RAN slice; Sum rate; Delay; Lyapunov optimization

I. INTRODUCTION

To improve the flexibility of services for 5G systems, wireless network virtualization (WNV) has emerged to abstract the physical network resources into virtual resources to realize sharing [1], [2]. Further, with WNV, end-to-end network slicing from the device to the core network could be created in order to satisfy users’ customized services [3]. Radio access network (RAN) slicing [4], as an important part of network slicing, specifically allows the virtual radio resources and virtual base stations (BSs) in RAN to be dynamically allocated to various services.

A lot of work has focused on the resource management and the analysis of its utility for RAN slices up to now. Specifically, the authors in [5] analyze the RAN slicing in a multi-cell network and its relationship to the radio resource management functionalities that can be used for splitting the radio resources among the RAN slices. And the authors in [6] consider a dynamic resource allocation criterion amongst tenants in RAN slices to optimize utility gains and capacity savings. However, there are still many other performance indicators for RAN slices remained to be investigated.

In particular, the ever-growing traffic demands and personalized user requests make the users’ sum data rate one of the major quality of service (QoS) goals for RAN slices. To ensure the fairness of the resources allocation in virtual networks, for service providers, [7] maximizes network throughput and reduces service time. A dynamic network slicing scheme for multitenant heterogeneous cloud radio access networks is proposed in [8] to achieve a higher network throughput. Furthermore, delay is viewed as a critical performance indicator related to the system reliability, which also reflects users’ QoS provided by RAN slices. A QoS model for a wireless virtualization network is developed in [9] to obtain the upper bound about network delay. [10] proposes the network slicing in fog radio access networks and formulates an average delay optimization problem for one slice with user data rate guarantee for the other slice. From another perspective, to achieve higher network efficiency, the joint analysis or tradeoff strike of performance indicators is very necessary in communication network. For instance, [11] and [12] study the performance tradeoff of the energy efficiency versus delay in WNV.

Despite existing research on RAN slices and the performance analysis in WNV, a holistic study on sum rate and delay tradeoff of RAN slices to better guarantee QoS is not available in the literature, which is the focus of this paper. Different from existing virtualization related works, our new RAN slice model is extended from single cell to multi-cell with randomly distributed BSs and user equipments (UEs) and the interference between cells is taken into account. Firstly, the optimization of the dynamic density of virtual BSs in RAN slices is investigated to maximize the sum rate while supporting users’ delay tolerance. Secondly, since this problem is challenging due to the absence of a closed-form solution for the users’ sum rate, we solve it by using the Lyapunov optimization, which reduces the implementation complexity effectively without loss of generality, because any prior knowledge of traffic rate is not required. Finally, the upper bounds of the delay and sum rate about user density, virtual BS density and a control factor are analytically derived to strike a tradeoff, and the impact of relevant parameters on the rate and delay performance is evaluated by numerical results.

II. SYSTEM MODEL

A. RAN Slices

We consider the RAN slices, as shown in Fig. 1. In this network, there is a mobile network operator (MNO). The buffer resources in BSs and radio spectrum resources in Air Interfaces owned by the MNO can be abstracted into multiple
virtual radio resources

Virtual buffer resources

Isolation

RAN slice 1

User 1

RAN slice 2

User 2

RAN slice K

User K

Figure 1. RAN slices.

Figure 2. The actual queueing process and virtual queue representations in RAN slices.

virtual resource units in the virtual resource pool. Besides, a physical BS will be virtualized into one or several virtual BSs, while MNO determines the amount of virtual buffer and radio resource units allocated to each virtual BS based on diverse location information and QoS requirements of users, further to form customized RAN slices with isolation guarantee to provide users’ access services. Once the user’s rate and delay requests, or channel conditions in the current networks change, the virtual buffer and radio resource allocation of RAN slices will be changed.

Without loss of generality, we assume a downlink wireless network scenario in which N physical BSs and multiple UEs are randomly distributed in the Euclidean plane with an area of \( S_{area} \). Therefore, we use the stochastic geometry here to model the BSs and UEs deployment. Through virtualization, we obtain specific RAN slices based on users’ location information and their customized QoS requirements. For simplicity, we suppose that each RAN slice is customized for one user. The system is operated time slotted and the slot is normalized to an integral unit described as a time interval \([t, t+1)\), \( t \in \{0, 1, 2, \ldots\} \).

As the virtual BSs and UEs distributions in RAN slices are time-varying due to the users’ mobility and the users’ customized requirements, the locations of virtual BSs and UEs are modeled by the dynamic homogeneous Poisson point process (PPP) \( \Phi_{VB} = \{ b_i, i = 0, 1, 2, \ldots \} \) with density \( \lambda_{VB}(t) \) and \( \Phi_U = \{ u_i, i = 0, 1, 2, \ldots \} \) with density \( \lambda_U(t) \), respectively, where \( t \in \{0, 1, 2, \ldots\} \). For simplification, it is assumed the virtual BSs and UEs deployment in our model is constant within a slot, and only changes from slot to slot. In addition, let vectors \( A(t) = \{ A_1(t), A_2(t), \ldots, A_u(t), \ldots \} \) and \( Q(t) = \{ Q_1(t), Q_2(t), \ldots, Q_u(t), \ldots \} \) denote the processes of random data arrivals and the current data queues in multiple RAN slices in slot \( t \), respectively, where \( A(t) \) is independent and identically distributed (i.i.d.) over time with the arrival rate \( \gamma \).

B. Rate Analysis of RAN slices

The maximum transmission power of a physical BS is denoted by \( P_B \), and the total power of the network is:

\[
P_{sum} = N \cdot P_B.
\]

(1)

\( P_C \) is used to indicate the circuit power consumption of a physical BS when there is no virtualization. Since virtualization consumes extra circuit power, the circuit power consumption of the virtual networks can be defined as:

\[
P_{VC}(t) = P_C \lambda_{VB}(t) S_{area}.
\]

(2)

When the density of virtual BSs increases, the circuit power consumption by virtualization will also increase. In addition, the maximum transmit power of each virtual BS is as follows:

\[
P_{VB}(t) = \frac{NP_B - P_C \lambda_{VB}(t) S_{area}}{\lambda_{VB}(t) S_{area}}.
\]

(3)

Then, we can define the time-average expectation of \( P_{VB}(t) \) as follows:

\[
\tilde{P}_{VB}(\lambda_{VB}(t)) = \lim_{t \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} E\{ P_{VB}(\tau) \}.
\]

(4)

Without loss of generality, each UE is associated with the closest virtual BS and \( r_{u,min} \) represents the distance between them. The power loss propagation is modeled by \( hr^{-\alpha} \), where the small-scale channel gain \( h \) follows an exponential distribution with unit mean. \( \alpha > 2 \) and \( r \) denote the path loss exponent and the propagation distance respectively. For a typical user \( u \), the signal to interference plus noise ratio (SINR) can be expressed as:

\[
\text{SINR}_u(t) = \frac{P_{VB}(t) h_{u,VB} r_{u,min}^{-\alpha}}{I_u(t) + \sigma^2}
\]

\[
= \frac{(NP_B - P_C \lambda_{VB}(t) S_{area}) h_{u,VB} r_{u,min}^{-\alpha}}{\lambda_{VB}(t) S_{area} (I_u(t) + \sigma^2)}, \forall u,
\]

(5)

where \( h_{u,VB} \) is the channel gain between user \( u \) and its nearest virtual BS \( VB \) and \( \sigma^2 \) is the noise power. \( I_u(t) \) is the interference from other virtual BSs, denoted as:

\[
I_u(t) = \sum_{\Phi_{VB,j} \neq i} P_{VB} h_{u,VB} r_{u,VB,j}^{-\alpha}
\]

\[
= \sum_{\Phi_{VB,j} \neq i} \frac{NP_B - P_C \lambda_{VB}(t) S_{area}}{\lambda_{VB}(t) S_{area}} h_{u,VB} r_{u,VB,j}^{-\alpha}, \forall u.
\]

(6)
C. Delay Analysis of RAN slices

At the same time, we can derive the time-average expectation the virtual network is as follows:

\[ R = \frac{1}{t} \int_0^t E\{R_u(t)\} \, dt, \]

and the pdf of the user’s distance \( r_{u,VB} \) from other interfering virtual BSs is:

\[ f(r_{u,VB}) = 2r_{u,VB}/R^2, \]

where \( R \) is the cell radius of the virtual network. We can find out the average value of interference and SINR received by user \( u \) at slot \( t \), respectively, given by

\[ I_u^v(t) = \int_0^\infty I_u(t)f(r_{u,VB})dr_{u,VB}f(r_{u,min})dr_{u,min}, \]

\[ = \int_0^\infty 2h_{u,VB}(NP_B - P_C \lambda_V B(t) S_{area})e^{-\lambda_V B(t)r^2_{u,min}} \cdot \frac{2\pi \lambda_V B(t)^2}{3r^2_{u,min}} \, dr_{u,min}, \]

and

\[ SINR_u^w(t) = \int_0^\infty SINR_u(t)f(r_{u,min})dr_{u,min}, \]

\[ = \int_0^\infty (NP_B - P_C \lambda_V B(t) S_{area}) h_{u,VB} \cdot \frac{e^{-\lambda_V B(t)r^2_{u,min}} \cdot 2\pi \lambda_V B(t)^2}{3r^2_{u,min}} \, dr_{u,min}, \]

According to the Shannon formula, the average data rate \( R_u(t) \) of the RAN slice can provide for user \( u \) at slot \( t \) is given as:

\[ R_u(t) = E[\log_2(1 + SINR_u^w(t))], \]

\[ \text{log}_2 \{ 1 + \int_0^\infty (NP_B - P_C \lambda_V B(t) S_{area}) h_{u,VB} \cdot \frac{e^{-\lambda_V B(t)r^2_{u,min}} \cdot 2\pi \lambda_V B(t)^2}{3r^2_{u,min}} \, dr_{u,min} \}, \]

Consequently, the downlink sum rate of all RAN slices in the virtual network is as follows:

\[ R_{sum}(\lambda_V B(t), \lambda_U(t)) = \lambda_U(t) S_{area} R_u(\lambda_V B(t)). \]

At the same time, we can derive the time-average expectation of the rate:

\[ R_{sum}(\lambda_V B(t), \lambda_U(t)) = \lim_{t \to \infty} \frac{1}{t} \int_0^t E\{R_u(t)\}. \]

C. Delay Analysis of RAN slices

To analyze the delay of RAN slices, based on Fig. 2, we model the queuing process of data requested by user \( u \) as

\[ Q_u(t + 1) = \max\{Q_u(t) - R_u(t), 0\} + \Lambda_u(t), \forall u. \]

Therefore, the average data queue length of all RAN slices in the network can be expressed as follows:

\[ Q = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \sum_{k=1}^{K(t)} E\{Q_k(t)\}, \]

where \( k \) denotes the \( k \)-th RAN slice and \( K(t) = \lambda_U(t) S_{area} \) denotes the total number of RAN slices in the network.

Using the Little Theorem [14], we can get the average delay of the network as

\[ D = \frac{Q}{\gamma}. \]

Without loss of generality, we use the length of the data queue to express the delay of the current network. Unless otherwise specified, the delay mentioned in this paper refers to the length of the data queue in the network.

In this paper, we study the problem of rate-delay tradeoff given the network is stable. A network is stable when all these discrete queues \( Q_k(t) \) are mean rate stable, that is, satisfying the following condition [15]:

\[ \lim_{t \to \infty} E\{|Q_k(t)\} = 0. \]

III. PROBLEM FORMULATION AND SOLUTION

According to (1), the power in the network is constant, so when the sum rate increases, the overall energy efficiency of the network increases. Therefore, in this section, we study the sum rate maximization of the RAN slices.

We develop a delay-constrained sum rate maximization problem. Mathematically, the problem can be formulated as

\[ \max \ R_{sum}(\lambda_V B(t), \lambda_U(t)) \]

s.t. \( C_1 : \bar{P}_V B \leq P_{av}^v, \forall k, \]
\( C_2 : Q_k(t) \leq \beta, \forall k, t, \]
\( C_3 : \lambda_V B(t) \geq 0, \forall t, \)
\( C_4 : P_{V B} \geq 0, \forall t, \)

where \( P_{av}^v \) and \( \beta \) denote the threshold values of the average power consumption by virtual BSs and the maximum tolerable delay requested by users, respectively. \( P_{av}^v \) is used to ensure that the physical BSs have sufficient energy to maintain and operate normally; and \( \beta \) is used to guarantee the users’ QoS. In addition, \( C_1 \) constraints the average power consumption by virtual BSs. According to (15), \( C_2 \) guarantees the stability of queues and limits the equivalent queue length in RAN slices. \( C_3 \) and \( C_4 \) give the non-negativity of the virtual BS density and the power.

Based on the general Lyapunov theory [15], we transform the average power constraint \( C_1 \) in (18) into a virtual power queue stable problem. The virtual power queue is defined as \( G_k(t) \), where \( G_k(0) = 0 \) and

\[ G_k(t + 1) = \max[G_k(t) + P_{V B}(t) - P_{av}^v, 0], \forall k, t. \]

Combining (14) and (15), we can rewritten the \( C_2 \) in (18) equivalently as the following constraint \( C_2' \):

\[ Q_k(t + 1) = \begin{cases} \max[Q_k(t) - R_k(t), 0], & \text{if } Q_k(t) > \beta, \\ \max[Q_k(t) - R_k(t), 0] + A_k(t), & \text{e.w.} \end{cases} \]
Considering (3), $C4$ can be equivalently transformed into the following constraint $C4$:

$$\lambda_{VB}(t) \leq \frac{NP_B}{P_C S_{\text{area}}}. \quad (21)$$

Then, (18) can be further recast equivalently to

$$\max \{\tilde{R}_{\text{sum}}(\lambda_{VB}(t), \lambda_U(t))\}$$

s.t. $C2, C3, C4$,

$$C1 : G_k(t) \text{ is mean rate stable}, \forall k. \quad (22)$$

To tackle $C1$ and $C2$, a combined vector, $\Theta(t) = [Q(t), G(t)]$, is defined to represent queuing states of all queues, where $Q(t)$ and $G(t)$ are virtual queues sets.

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

$$\Delta[\Theta(t)] = E\{L[\Theta(t+1)] - L[\Theta(t)] | \Theta(t)\}, \quad (23)$$

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

$$L[\Theta(t)] = \frac{1}{2} \sum_{k=1}^{K(t)} Q_k(t)^2 + \frac{1}{2} \sum_{k=1}^{K(t)} G_k(t)^2. \quad (24)$$

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

$$F(t) = \Delta[\Theta(t)] - VE[R_{\text{sum}}(t)], \quad (25)$$

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

We have the following inequality about $F(t)$ by using Lyapunov optimization theory [15],

$$F(t) \leq B + \sum_{k=1}^{K(t)} \xi[Q_k(t) - \beta]Q_k(t)E[A_k(t)\Theta(t)]$$

$$+ \sum_{k=1}^{K(t)} G_k(t)E[P_{VB}(t) - P_{VB}^\text{av}]\Theta(t)$$

$$- \sum_{k=1}^{K(t)} Q_k(t)E[R_k(t)\Theta(t)] - VE[R_{\text{sum}}(t)\Theta(t)], \quad (26)$$

where $B \geq 0$ is a limited constant and satisfies the following condition:

$$B \geq \frac{1}{2} \sum_{k=1}^{K(t)} E\{[\xi[Q_k(t) - \beta]A_k(t)]^2 + R_k(t)^2\Theta(t)\}$$

$$+ \frac{1}{2} \sum_{k=1}^{K(t)} E\{(P_{VB}(t) - P_{VB}^\text{av})^2\Theta(t)\}, \quad (27)$$

where $\xi$ is a function of $Q_k(t) - \beta$ and satisfies: $Q_k(t) - \beta \leq 0$, $\xi[Q_k(t) - \beta] = 1$; otherwise, $\xi[Q_k(t) - \beta] = 0$, which satisfies $C2$ in (22).

Specifically, with the help of the stochastic optimization theory [15], the optimal solution of the above problem (22) can be obtained by minimizing the upper bound of $F(t)$ slot by slot, i.e., solving the problem (28) below:

$$\min_{k=1}^{K(t)} \{G_k(t)[P_{VB}(t) - P_{VB}^\text{opt}] - Q_k(t)R_k(t) + \xi[Q_k(t) - \beta]Q_k(t)A_k(t)\} - V[R_{\text{sum}}(t)]$$

s.t. $C3 : \lambda_{VB}(t) \geq 0, \forall t$,

$$C4 : \lambda_{VB}(t) \leq \frac{NP_B}{P_C S_{\text{area}}}, \forall t. \quad (28)$$

At this point, we have transformed the original difficult problem (18) into the problem (28) that is easy to solve.

Therefore, we can propose an effective virtual resource allocation algorithm based on the classic drift-plus-penalty algorithm, summarized in Algorithm 1. In particular, the solution to equation (28) can be found using a simple one-dimensional search.

**Algorithm 1 Dynamic virtual resource allocation algorithm to solve (28).**

1: Initialization: $Q_k(0)=0$, $G_k(0)=0$, and $R_{\text{sum}}(0)=0, \forall k$.
2: Repeat:
3: Update virtual resource allocation $\lambda_{VB}(t)$ according to (28).
4: Let $t = t + 1$.
5: Update $Q_k(t)$, $G_k(t)$ and $R_{\text{sum}}(t)$ according to (20), (19) and (12), respectively.
6: Stop when $t = T$, where $T$ is the total number of time slots.

**IV. TRADEOFF BETWEEN RATE AND DELAY**

In this section, the upper bounds of the sum rate and delay, as well as their tradeoff will be derived.

**Theorem 1:** The sum rate in (18) is bounded by

$$R_{\text{sum}}^\text{opt} \geq \tilde{R}_{\text{sum}}(\lambda_{VB}(t), \lambda_U(t)) \geq R_{\text{sum}}^\text{opt} - \frac{B}{V}. \quad (29)$$

**Proof:** Assume that $R_k^*(t)$, $P_{VB}^*(t)$ and $R_{\text{sum}}^*(t)$ result from an feasible virtual BS allocation policy $\lambda_{VB}(t)$ of problem (18) in slot $t$. The following conditions for any $\delta > 0$ and $\varepsilon > 0$ can be obtained according to the stochastic optimization theory [15]:

$$E[A_k(t) - R_k^*(t)|\Theta(t)] = E[A_k(t) - R_k^*(t)] \leq \delta, \quad (30)$$

$$E[R_{\text{sum}}^*(t)|\Theta(t)] = E[R_{\text{sum}}^*(t)] \geq R_{\text{sum}}^\text{opt} + \delta, \quad (31)$$

$$E[R_k^*(t)|\Theta(t)] = E[R_k^*(t)] \geq \gamma_k + \varepsilon, \quad (32)$$

$$E[P_{VB}(t) - P_{VB}^*(t)|\Theta(t)] = E[P_{VB}(t) - P_{VB}^*(t)] \leq \delta. \quad (33)$$

where $\gamma_k$ is the data arrival rate of each RAN slice.

Substituting (30)-(33) into the right-hand side of (26) and letting $\delta \to 0$, we have the following inequality:

$$\Delta[\Theta(t)] - VE[R_{\text{sum}}(t)|\Theta_t]$$

$$\leq B + \sum_{k=1}^{K(t)} \{\xi[Q_k(t) - \beta] - 1\}Q_k(t)\gamma_k$$

$$- \varepsilon \sum_{k=1}^{K(t)} Q_k(t) - VR_{\text{sum}}^\text{opt}(t). \quad (34)$$
Taking expectation for both sides of (34), and then using telescope sum over \( t \in \{0, 1, ..., H - 1\} \) for the result as well as taking the fact that \( Q_k(t) \geq 0 \) and \( \xi(Q_k(t) - \beta) - 1 \leq 0 \) into consideration, we have
\[
E\{L[\Theta(H)]\} = E\{L[\Theta(0)]\} - V \sum_{t=0}^{H-1} E[R_{sum}(t)] \\
\leq HB - HV R_{sum}^{opt},
\]
where \( H \) represents the number of the time slots.

Dividing both sides of the above inequality by \( VH \) yields the following:
\[
-\frac{E\{L[\Theta(0)]\}}{VH} - \frac{1}{H} \sum_{t=0}^{H-1} E[R_{sum}(t)] \leq \frac{B}{V} - R_{sum}^{opt} \tag{36}
\]
Let \( H \to \infty \) and consider equation \( E\{L[\Theta(0)]\} \geq 0 \), we can obtain the following conclusion
\[
R_{sum}(\lambda_{VB}(t), \lambda_U(t)) = \frac{1}{H} \sum_{t=0}^{H-1} E[R_{sum}(t)] \\
\geq R_{sum}^{opt} - \frac{B}{V}.
\]
Thus, the (29) can be proved.

**Theorem 2:** The average queue length is bounded by
\[
\bar{Q} = \lim_{H \to \infty} \frac{1}{H} \sum_{t=0}^{H-1} \left[ \sum_{k=1}^{K(t)} E\{Q_k(t)\} \right] \\
\leq \min\{\lambda_u S_{area} \beta, \frac{B + V(\lambda_u S_{area} R_{kmax}^{opt} - R_{sum}^{opt})}{\varepsilon}\},
\]
where \( R_{kmax}^{opt} \) is the rate threshold available for a RAN slice, and \( \lambda_u \) is the time-average expectation of \( \lambda_U(t) \).

**Proof:** When \( Q_k(t) \leq \beta \), take expectation for (34) while using telescoping sum over \( t \in \{0, 1, ..., H - 1\} \) for the result. Considering \( \xi(Q_k(t) - \beta) = 1 \), we have
\[
E\{L[\Theta(H)]\} - E\{L[\Theta(0)]\} - V \sum_{t=0}^{H-1} E[R_{sum}(t)] \\
\leq HB - HV R_{sum}^{opt} - \varepsilon \sum_{t=0}^{H-1} \sum_{k=1}^{K(t)} E[Q_k(t)].
\]
Dividing both sides of the above inequality by \( \varepsilon H \) and taking \( H \to \infty \) we have:
\[
\bar{Q} = \lim_{H \to \infty} \frac{1}{H} \sum_{t=0}^{H-1} \left[ \sum_{k=1}^{K(t)} E[Q_k(t)] \right] \\
\leq \frac{B - V R_{sum}^{opt}}{\varepsilon} + \frac{V}{\varepsilon} \lim_{H \to \infty} \frac{1}{H} \sum_{t=0}^{H-1} E[R_{sum}(t)],
\]
where
\[
\lim_{H \to \infty} \frac{1}{H} \sum_{t=0}^{H-1} E[R_{sum}(t)] \leq R_{sum}^{max}, \tag{41}
\]

\[
\bar{Q} = \lim_{H \to \infty} \frac{1}{H} \sum_{t=0}^{H-1} \left[ \sum_{k=1}^{K(t)} E[Q_k(t)] \right] \\
\leq \frac{B - V R_{sum}^{opt}}{\varepsilon} + \frac{V}{\varepsilon} \lim_{H \to \infty} \frac{1}{H} \sum_{t=0}^{H-1} E[R_{sum}(t)],
\]
Furthermore, we can easily obtain the following conclusion when \( Q_k(t) \geq \beta \), or
\[
\bar{Q} = \lambda_u S_{area} \beta. \tag{43}
\]
Combining both (42)and (43) completes the proof.

**V. NUMERICAL RESULTS AND DISCUSSIONS**

In this section, extensive numerical evaluation results are provided to validate the theoretical findings. The transmit power of a BS is assumed to be \( P_B = 20W \) and the number of physical BSs is \( N = 2 \). In addition, we set \( \alpha = 4 \), \( \lambda_u = 0.000125 \text{ (bits/slot/Hz)} \) and \( S_{area} = 2km \times 2km \). Moreover, we use 10,000 slots to approximate \( t \to \infty \).

Fig. 3 indicates the network performance versus data arrival rate \( \gamma \) with different settings, where \( V \) is the control factor of the tradeoff between the sum rate and delay. The average delay of one RAN slice increases exponentially with \( \gamma \) at first and then tends to be stable when the network queueing length exceeds the limit of \( \beta \). Similarly, the sum rate decreases with \( \gamma \) and eventually converges because of the limit of \( P_{VB} \) and \( \beta \).

Fig. 4 illustrates the impact of \( V \) on the sum rate and delay for RAN slices under different \( \gamma \). The sum rate tends to converge with the increase of \( V \). When \( V \) approaches infinity, the sum rate converges to the optimal rate of the system. Then the average delay of one RAN slice shows a linear growth trend with the increase of \( V \) because the queue length does not exceed the limit of \( \beta \). In addition, we can see that \( \gamma \) has a negative impact on the network performance, that is, when \( V \) is the same, the larger \( \gamma \), the lower the sum rate as well.
as the larger the average delay. Finally, as we can see, the optimization of the sum rate can be achieved at the cost of the delay. The lower the delay, the lower the sum rate. There is a clear performance trade-off between the two metrics, as discussed in the previous section.

Fig. 5 describes virtual BS density versus data arrival rate with different settings of $V$. We can see that the optimal virtual BS density increases at the beginning, and then tend to be stable because of the limit of $P_{V_B}^{av}$ and $\beta$ with $\gamma$. That is, when the data arrival rate in the network increases, we can increase system performance by increasing the number of virtual BS properly in the system. In addition, $V$ has a positive impact on network performance. The network can transmit larger amount of data for users in the same virtual BS deployment with increased of $V$.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, the rate-delay tradeoff problem is investigated in RAN slices with randomly distributed BSs and UEs. The problem is formulated as a stochastic optimization problem to optimize the sum data rate with the constraint on the delay. Afterwards, it is solved using the Lyapunov optimization. Finally, the theoretical bounds for both the sum rate and delay are derived, which reveal the way of adjusting the parameter $V$ to get a performance tradeoff. Future work is in progress to consider caching and computing resource allocation in RAN slices.

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (61771358), National Natural Science Foundation of Shaanxi Province (2018JM6052), Inter-governmental International Cooperation on Science and Technology Innovation (2016YFE0122900), and the 111 Project (B08038).

REFERENCES