A Random Compensation Scheme for 5G Slicing Under Statistical Delay-QoS Constraints

XUEFEN CHI, YIHANG JING, HONGLIANG SUN, AND BAOZHU YU

1Department of Communications Engineering, Jilin University, Changchun 130012, China
2College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China

Corresponding author: Xuefen Chi (chixf@jlu.edu.cn)

This work was supported in part by the Natural Science Foundation of the Department of Science and Technology of Jilin Province under Grant 20180101040JC, and in part by the National Natural Science Foundation of China under Grant 61801191.

ABSTRACT For the communication scenarios of heterogeneous arrivals in the fifth generation cellular networks, slicing is believed to be an effective way to provide quality-of-service (QoS) provisioning. However, the reports on an effective segmentation of resources for random and bursty arrival are rare. Relying on the bandwidth estimation and service design, we give a simple but effective scheme for the slicing of bandwidths named two-hole leaky bucket random service mechanism. The service in our scheme comprises two parts, the specific basic service and shared random compensation service. Accordingly, the mechanism of resource allocation is just like leaky bucket with two holes, one of which has a random leak rate. To guarantee the statistical QoS of heterogeneous traffic flows, effective bandwidth/capacity theory is introduced into the slicing scheme. By establishing function equations, we deduce the compensation probability of each slice, which is taken as the slicing adjustment factor. Simulation results prove that the random compensation scheme can satisfy the statistical QoS requirements of heterogeneous traffic flows. Although the scheme is proved to be effective against two slices, the proposed model is applicable to general multiple heterogeneous arrivals.

INDEX TERMS Quality of service, network slicing, resource allocation, effective bandwidth, effective capacity.

I. INTRODUCTION

The fifth generation (5G) cellular networks promise to carry a greater number of terminals and support more types of services. Three use case categories with stringent heterogeneity are already classified as enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive internet of things (MIoT). The high quality video streaming and fast large file transmission are two kinds of typical eMBB applications. The URLLC is urgently needed by remote control systems and industrial automation. The MIoT scenario contains a large number of devices with high density in space. These services possess different quality of service (QoS) requirements. The network slicing has been considered as the most effective 5G core technology to provide customized QoS provisioning for heterogeneous services. The basic concepts related to Network Slice (NS) have been defined in 3GPP TS-23.501 [1]. The network slices may differ from supported features of the Slice/Service Type and network functions optimization.

Orienting a new feature of the committed service, delivering the dedicated 5G QoS flow requires to deploy a separate network slice. In terms of the objective functions for QoS provisioning optimization, to maximize slice service rate and/or minimize delay, the slices are classified as follows: delay constrained slice, rate constrained slice, delay and rate constrained slice, and delay and rate non-constrained slice [2]. In the actual slicing scenarios, the allocation of resource blocks (RBs) could be implemented by software-defined networking and virtualized network functions [3].

The stochastic feature of versatile traffic arrivals in 5G makes network slicing with QoS requirements a challenging but also an interesting task. The scheduling approaches to orchestrate slices have been widely explored for reachable transmission rate, reliability, and interrupted transmission rate [4]–[6]. However, the delay, as an important QoS metric was not evaluated in their works.

As the characteristics of arrivals change in time and the radio channel conditions are time-varying, it is challenging to slice the bandwidth resources efficiently and accurately. As a result, only a few bandwidth slicing algorithms are reported up to now. Besides, almost all the reported resource
slicing schemes are immersed in the scheduling process now and they are realized by continuously monitoring channel conditions and then allocating resource blocks, in some cases which are unnecessarily complicated. Our previous research has indicated that the bandwidth cost by two aggregated random arrivals is less than the sum of bandwidths cost by two individual random arrivals. This suggests that bandwidth sharing between traffic flows could contribute radio resource management. However, from another view angle, isolation between slices is a basic thinking in slicing. Targeting to use bandwidth effectively and keeping slices in an appropriate isolation, a two-hole leaky bucket (LB) random service mechanism is explored for the bandwidth slicing in this paper. Inspired by the work of [7], the service mechanism is depicted as a bucket with two leaky holes. The leak rates of the holes are corresponding to the basic and random compensation services rates respectively. In our scheme, every slice provides basic service using its dedicated bandwidth, and two or more slices share a bandwidth to provide compensation service. The scheme we designed operates in a random compensation way to avoid frequent resource monitoring and meticulous resource allocation. The queuing theory is adopted to model the system. The effective bandwidth/capacity (EB/EC) theory is employed to realize our method and formulate the random compensation service.

We aim to propose a simple but effective slicing scheme and provide customized QoS provisioning for the heterogeneous traffic flows. The random compensation probabilities of slices, which are solved by the equations established from EB/EC theory, are defined as slicing adjustment factors. The goal of the paper is to get slicing adjustment factors and obtain the random compensation bandwidths for slices. The contributions can be summarized as follows:

- We propose a new random compensation service scheme for network slice instances, which approves the statistical characteristics of arrival flows and fading channel. We define the slicing adjustment factor to guide the dynamic allocation of bandwidth resources.
- Effective bandwidth/capacity theory is introduced into the two-hole LB random service mechanism for the first time, which helps to guarantee the QoS requirements of heterogeneous services. The compensation probability and compensation bandwidth are both studied and are developed as two new independent variables in effective capacity function.
- The simulations indicate: (a) our scheme supports strong heterogeneous traffic flows effectively and as a result it is particularly suitable for mixed service such as sustained used video call; (b) for services with Poisson-like arrival (such as voice), the random compensation scheme presents specially satisfied performance. It is verified by simulations that less compensation bandwidth could support wide range QoS constraints.

The rest of this paper is organized as follows: the related works are reviewed in Section II. The system models are established in Section III. In Section IV, compensation bandwidths and compensation probabilities deduced by EB/EC are solved. Simulation scenarios and evaluation results of the proposed method are demonstrated in Section V. Finally, the conclusion of this work is given in Section VI.

II. RELATED WORKS

The slicing schemes with QoS provisioning have been studied in wireless networks. In [2], a virtual resource slicing scheme was proposed for the next generation radio access network (RAN). The reinforcement learning technique was introduced to resource allocation by reserving unused resources dynamically. In [3], the authors studied autonomous resource provisioning and customization in RAN. Bennis et al. analyzed downlink resource slicing for 5G new applications in [4]. The problem of punctured scheduling eMBB and URLLC was constructed as a risk-sensitive optimization problem. To guarantee the heterogeneous QoS of traffic flows, the authors of [5] investigated the heterogeneous non-orthogonal multiple access (H-NOMA). In the H-NOMA slicing scheme, the transmission rate and the supportable average arrival rate of eMBB, URLLC, and MiToW were studied for single cell. In collaboration with Simeone et al., they extended the research to multicell fog RANs in [6]. In [7], an extended token bucket fair queuing (e-TBFQ) scheduling method was proposed for uplink spectrum slicing. The e-TBFQ scheme scheduled tokens among an LB and two levels of token buckets for each application flow. In [8], the authors figured out the impact of network slicing on RAN requirements and designed the templates for 5G RAN slicing scheme. In [9], a network slicing game was formulated to maximize the utility of a single user. In cloud RANs, Jiang et al. [10] used an auction approach to model the 5G slicing scheme under the satisfaction requirements. In [11], similar to the OpenFlow protocol, a network slicing matching scheme was designed based on the proposed recursive architecture. A slicing mechanism of resource abstraction and link rate instantiation was proposed [12], in which the radio resource was abstracted reasonably so as to hide the complex details of fading channel. Saad et al. used M/M/1 queuing model to abstract the base station-side buffer queue in [13]. The transmission power and time-frequency resource allocation were optimized to meet the delay constraints on every slice. Meugen et al. studied the resource allocation problem of eMBB and URLLC slices based on Lyapunov optimization theory in [14]. The Markov’s inequality was used to derive the queue overflow probability of arrival flows. The comparisons of different approaches are listed in TABLE 1. Based on our knowledge and literature review, few of schemes in [2]–[14] considered the heterogeneity of the traffic. The slicing schemes of [2]–[11] have explored reachable transmission rate, interrupted transmission rate, and many other QoS parameters. However, the statistical delay-QoS, as an important QoS metric was not evaluated in their works. The random features of arrival and service processes should be studied for soft QoS requirements. Hence, one of the main study points in our paper is whether the statistical delay-QoS analysis method...
has been introduced. The comparison criteria of realization strategy and complexity have also been considered in Table 1. The existing slicing orchestration algorithms were commonly immersed in the design of scheduling processes. In such method, it was unnecessarily complicated under the statistical QoS constraints. For example, most of the slicing solutions adopted the strategy of continuously monitoring resource blocks and then allocating them. In our paper, a new network slice pattern defined as the random compensation slice has been designed to cope with the random factors of communications. Both effective bandwidth and effective capacity has been introduced to our scheme for studying heterogeneous slices.

| Criterion Scheme | Traffic heterogeneity | Statistical delay-QoS analysis | Strategy of realization | Difficulty of realization |
|-------------------|-----------------------|-------------------------------|-------------------------|--------------------------|
| Schemes of [2] - [11] | [4], [7] - [11] unconsidered | None | Monitored and scheduled every RB | Complicated |
| Schemes of [12] - [14] | [12], [14] unconsidered | Yes | Randomly allocate a RB set for slices | Simple |
| Our scheme | Considered | Yes | |

Effective bandwidth (EB) referred to the required constant service rate for the traffic under given statistical delay-QoS requirements [15]. The authors in [16] estimated the EB of the self-similar arrival. The authors in [17] derived the EB of the arrivals with bursty characteristic in ultra-reliable and low-latency RANs. Effective capacity (EC) is a newly emerging delay-QoS guarantee theory in recent years. The Shannon capacity was applied to derive the EC in [18]. EC referred to the transmission rate supported by the wireless link under given statistical QoS constraints [19]. In [20], Tang et al. carried out the research on delay-QoS in mobile wireless communication systems. In addition, EC was also studied to guarantee QoS in the scenarios like cross-layer design of finite block length coding over 5G mobile wireless networks [21], [22] and the random access system of visible light communications [23], [24]. In [25], the authors used the EC to characterize the user experience rate of eMBB. The scheduling algorithm with the system capacity was proposed. Realizing the long-term statistical QoS provisioning is not an easy task. The authors in [26] proved that the statistical QoS could be guaranteed when EC approaches EB. In [12], EC was used to characterize the wireless link service rate while EB was used to characterize the bandwidth requirements of QoS constrained services. EB/EC theory provides an important reference to the statistical QoS guarantees.

III. SYSTEM MODEL

A. VIRTUALIZATION NETWORK MODEL

The relation between the essential concepts of the slicing scheme is manifested in Fig. 1. In this paper, network slice instance (NSI) is a service instance for the traffic flows which are mapped to the same 5G QoS flow. An NSI is identified by Network Slice Selection Assistance Information (NSSAI), and NSI may contain several network slices. A network slice is identified by a unique Single Network Slice Selection Assistance Information (S-NSSAI). NSSAI is the set of S-NSSAIs. S-NSSAI can reflect the service ability of the corresponding networking slice. For example, the first S-NSSAI in NSSAI may determine whether the slice can serve the bursty traffic well, and the second S-NSSAI may indicate the statistical delay-QoS level that can be guaranteed by the NSI. In our scheme, an admitted flow is served by a network slice instance, which consists of a basic service network slice and a shared random compensation service network slice.
In our study, delay-QoS refers to statistical QoS, which has two necessary parameters, delay and violation probability of QoS. It is the appropriate service rate that guarantees delay-QoS. In terms of NSI 1, the model of MMOO traffic has two states. In On state, the number of arrival packets at one slot is $R$, and the number of arrival packets in Off state is zero. The transition matrix of MMOO arrival process $T$ is

$$T = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

(2)

where the transition probability from Off to On state is $p$, and the one from On to Off state is $q$. The mean arrival rate of MMOO process is $R_{av} = pR/(p+q)$. The burstiness of MMOO arrival could be measured by $c^2 = q(2p-q)/((p+q)^2)$, where $c^2$ is the variance coefficient [28]. In NSI 2, the expectation of the Poisson arrival process is represented as $\lambda$.

The random compensation service resources can be multiplexed by NSIs. In each time slot, compensation slice serves at least one traffic flow. The fading channel is modeled as the random blocked channel [29]. The shortage of bandwidth caused by the random fading channel state could be countered by the random compensation service process. The random compensation service NS 0 is configured for the target NSI $j$ to satisfy the statistical QoS. With this solution, the service process can serve the traffic without sending pilot frequently and save bandwidth resources.

The service process, $s_j(n)$, is modeled as a special Bernoulli process [30].

$$s_j(n) = \begin{cases} r_j + C & P_j \\ r_j & 1 - P_j \end{cases}$$

(3)

where $r_j$ and $C$ represent basic service rate and compensation service rate respectively. Benefiting from the ID of arrival traffic, basic service rate is set as the minimum of effective bandwidth, i.e., the effective bandwidth without considering QoS. $C$ is the compensation bandwidth sliced to network slice instances. The random service compensation probability $P_j$ is taken as the slicing adjustment factor that will be studied in the next section. The accumulated service process is represented as $S_j(n)$.

### IV. PROBLEM FORMULATION

#### A. RESEARCH METHODOLOGY

5G slicing technology in Fig. 1 realizes dynamical resource allocation and performance isolation among network slice instances. The existing slicing schemes scheduled or reserved resources by frequent resource monitoring and meticulous resource allocation. However, such slicing orchestration algorithms were usually unnecessarily complicated because they immersed in the scheduling processes, which were difficult to be abstracted as random service processes. Hence, we believe that the current slicing methods may have the drawback in bandwidth utilization. In this paper, a new random compensation slice pattern defined as random compensation slice

![FIGURE 3. Compensated service scheme for bandwidth resource slicing.](image)

Classified flows are served by NSI $i$ (LB $i$) which comprises specific NS $i$ (normally open hole) and shared NS 0 (random leaking hole controlled by switch $P$). The sizes of two holes are $r$ and $C$ respectively.
is designed to cope with the random factors in communications. We adopt the stochastic methods to characterize the performance of the slicing system model under statistical delay-QoS constraints. The scheme introduces EB/EC theory by which the EBs of the random arrival processes and the ECs of the designed service processes are derived. By formulating the relationship between EB and EC pairs, the slicing adjustment factors and random compensation service demanded by the heterogeneous traffic could be deduced. By formulating the relationship between EB and EC, the service rate demanded by the traffic could be deduced. The basic service rate could be obtained by considering the arrival traffic characteristics only, while combined with basic service rate, the compensation service rate undertakes the responsibility of statistical QoS provisioning. By using the random compensation slice with an appropriate probability, the statistical delay-QoS could be guaranteed and further bandwidth utilization could be improved. Therefore, the main solutions of the work are to design the algorithm to find a common compensation service rate and then to determine the probabilities with which every NSI uses random compensation slice.

**B. ANALYSIS OF PROBABILITY INEQUALITIES**

The instantaneous length of queue in NSI $j$ is defined as

$$Q_j(n) = \sum_{k=0}^{n} (a_j(k) - s_j(k))^+$$  \hspace{1cm} (4)

where $(X)^+$ operation returns the larger one between zero and $X$. The QoS exponent $\theta_j$ can reflect the delay constraints on the corresponding NSI. Based on the Large Deviation Theory, $\theta_j$ is given as

$$\theta_j = - \lim_{x_j \to \infty} \frac{\ln (P\{Q_j(\infty) > x_j\})}{x_j}$$  \hspace{1cm} (5)

where $Q_j(\infty)$ is the queue length in steady state and $x_j (x_j \geq 0)$ is the threshold of the queueing backlog in the NSI $j$.

Based on the condition that system queue can step into the steady state, queue overflow probability inequality is

$$P(Q_j(\infty) \geq x_j) \leq e^{-x_j \theta_j} \triangleq \varepsilon$$,  \hspace{1cm} (6)

where $\varepsilon$ is queue overflow probability threshold. The delay at slot $n$ and arrival intensity on NSI $j$ are represented as $D_j(n)$ and $\lambda_j$ respectively. Relying on the Little’s Law [31], $D_j(n) = Q_j(n)/\lambda_j$, the delay-violation probability inequality can be given as

$$P(D_j(\infty) \geq d_j) \leq e^{-\lambda_j d_j \theta_j} \triangleq \varepsilon$$,  \hspace{1cm} (7)

where $d_j$ can be regarded as the delay constraint of the flow in NSI $j$ and $\varepsilon$ is the delay-violation probability threshold.

**C. EFFECTIVE BANDWIDTH AND EFFECTIVE CAPACITY EQUATIONS**

Based on the concept of QoS exponent, the EB functions for the MMOO and Poisson arrival processes, $E_{B, MMOO}(\theta_1)$ and $E_{B, Poisson}(\theta_2)$ can be described as

$$E_{B, MMOO}(\theta_1) = \ln sp\left(\frac{e^{\theta_1 T}}{\theta_1}\right) - \frac{1}{\theta_1} \ln \left(1 - p + (1 - q)e^{\theta_1}/2\right) + \sqrt{(1 - p + (1 - q)e^{\theta_1})^2 - 4(1 - p - q)e^{\theta_1}}$$ \hspace{1cm} (8)

$$E_{B, Poisson}(\theta_2) = \ln E\left[\frac{e^{\theta_2 A_2(n)}}{\theta_2}\right] = \frac{\lambda e^{\theta_2} - 1}{\theta_2}$$, \hspace{1cm} (9)

where formula (9) is clearly presented in [27] and $sp(\cdot)$ is the spectrum radius of the matrix. $\Delta$ is the diagonal rate matrix of MMOO, which is given as

$$\Delta = \begin{bmatrix} 0 & \cdot \\
R & \ddots \end{bmatrix}$$ \hspace{1cm} (10)

The EC function, which indicates the bandwidth resources allocated to the corresponding NSI, can be expressed as follows

$$E_{C, j}(\theta_j, P_j, C) = -\frac{\ln E\left[\frac{e^{-\theta_j S_j(n)}}}{\theta_j}\right]}{\theta_j} = -\frac{\ln (P_j e^{-C\theta_j} + 1 - P_j)}{\theta_j} + r_j$$, \hspace{1cm} (11)

where $P_j$ and $C$ are new independent variables of the effective capacity formula in this paper.
The QoS exponent $\theta_j$ influences the values of EB and EC in opposite trend [29]. According to the arrival and random service models we proposed, we enable two network slice instances in the system and enumerate following relations

NSI 1: $E_{B,\text{MMOO}}(\theta_1) = E_{C,\text{1}}(\theta_1, P_1, C)$,

NSI 2: $E_{B,\text{Poisson}}(\theta_2) = E_{C,\text{2}}(\theta_2, P_2, C)$,

where $E_{C,\text{1}}(\theta_1, P_1, C)$ and $E_{C,\text{2}}(\theta_2, P_2, C)$ is the EC functions for the service processes in NSI 1 and NSI 2 respectively. $P_1 + P_2 = 1$ is considered as a constraint. The equations are established to acquire compensation probability $P_j$, $j = 1, 2$ and compensation bandwidth $C$ for the NSIs. Referring to factors $P_j$ and $C$, the division scheme for network slices can be adjusted dynamically. If the system bears other kinds of arrival traffic in arbitrary NSI $i (i = 1, 2, \ldots)$, the equation of $E_{B,\text{flow}}(\theta_i) = E_{C,i}(\theta_i, P_i, C)$ can be also established to find out $P_i$ for NSI $i$. The EC functions are also subject to $\sum P_i = 1$.

### V. SIMULATION RESULTS

The effectiveness of the random compensation scheme is investigated firstly and then performance of our scheme is evaluated. The parameters used in the following simulations are listed in Table 2. To make the comparisons easy and clear, we study the case where two network slice instances (NSI 1 and NSI 2) carry different application flows. All simulations are implemented on MATLAB.

#### A. SCHEME VALIDATION

We first investigate the availability of our proposed random slicing scheme (depicted in Fig. 2). In this simulation, two arrivals of the queuing systems are MMOO and Poisson processes. The average arrival rates of two arrival flows are same to be $\lambda = R_{av} = 10$ packets/slot. For MMOO arrival, the state transition probabilities (Off to On and On to Off) are $p = q = 0.3$, and so the arrival rate at state On is $R = R_{av}(p + q)/p = 20$ packets/slot. The burstiness of MMOO is measured by $c^2 (= q(2−p−q)/(p + q)^2 = 1.17)$. Delay-QoS constraint of Poisson traffic on NSI 1 is fixed to be 20 time slots. The delay constraint of MMOO traffic on NSI 1 is varying, and takes values from 0 to 160 time slots. The threshold of queue overflow probability $\varepsilon'$ is set to be $10^{-3}$ for both MMOO and Poisson arrivals. The simulations span $10^5$ time slots and are done 20 times. The results are illustrated in Fig. 4 (a)-(d) as box-plots form. The simulation results match the analytical results. The slicing adjustment factor $P_j$ varies with different delay constraints and the types of arrivals. And the queue overflow probabilities are almost all below corresponding threshold ($10^{-3}$), which indicates that the random compensation slicing scheme works well. In [3], the queue length was not compared with the length threshold to obtain the statistical queue overflow probability. This kind of hard guarantee may lead to excessive bandwidth consumption. Besides, their results were about the simulations of the certain specific traffic flow in a business scenario, while the real application scenarios of 5G services are complex and changeable. In this part, the queue overflow probability of

![FIGURE 4. Queue overflow probabilities of MMOO and Poisson traffic flows under delay constraints. The queue overflow probability threshold is $10^{-3}$. The delay constraints of MMOO flow are variable while the delay constraint of Poisson is 20 time slots ($c^2 = 1.17, R_{av} = \lambda = 10$ packets/slot). (a, b) NSI 1 for MMOO flow (c, d) NSI 2 for Poisson flow.](image-url)
the proposed scheme has been validated. The traffic with different characteristics will be studied in the performance evaluation part in which the diversified 5G businesses are served by the random compensation scheme.

B. PERFORMANCE EVALUATION

1) ARRIVAL INTENSITY
In this simulation, we evaluate the performance of the scheme in terms of arrival intensity. Our scheme adjusts compensation bandwidth and probability to meet QoS demands of two traffic flows, as shown in Fig 5. Delay-QoS constraint of Poisson traffic is fixed to be 20 time slots. The delay constraint of MMOO flow is 10 packets/slot. For MMOO flow, it is set to be \( R_{av} = \{5, 10, 15\} \) packets/slot respectively. For Poisson flow, average arrival rate \( \lambda = 10 \) packets/slot. The results show that MMOO flow costs compensation bandwidth with the obviously greater probability. The arrival intensity of MMOO flow influences compensation probability even in the cases that delay restrictions get loose. However, comparing to compensation probability, compensation bandwidth is not sensitive to the arrival intensity and the delay requirements when its delay demands get loose, and also under these cases small bandwidth compensates two strong heterogeneous traffic flows perfectly.

\[ P( R_{av} = 5) \]
\[ P( R_{av} = 10) \]
\[ P( R_{av} = 15) \]
\[ C( R_{av} = 5) \]
\[ C( R_{av} = 10) \]
\[ C( R_{av} = 15) \]

FIGURE 5. The relationships among MMOO arrival intensity, delay constraints of MMOO flow, compensation probability, and compensation bandwidth in NSI 1 (\( \epsilon \) for Poisson flow is 10^{-3}, \( \epsilon \) for MMOO flow is 10^{-3}, \( c^2 = 1.17 \), \( \lambda = 10 \) packets/slot). The delay constraint of Poisson in NSI 2 is 20 time slots.

2) BURSTINESS OF TRAFFIC
In this simulation, we evaluate the scheme performance based on burst traffic. The burstiness of traffic could be measured by variance coefficient ( \( c^2 \)). Fig. 6 depicts the impact of \( c^2 \) on compensation probability \( P \) and compensation bandwidth \( C \). In our simulations, the settings of delay constraints and \( \epsilon \) are same as the setting in Fig. 5. Average arrival rates of two arrival flows are same to be \( \lambda = R_{av} = 10 \) packets/slot. Adjusting state transition probability of MMOO arrival (let \( p = q = 0.25, p = q = 0.2, p = q = 0.167 \) respectively), the burstiness of MMOO is to be \( c^2 = 1.5, 2.0, 2.5 \) accordingly. It can be found that burstiness of traffic influences compensation scheme greatly. The traffic with bigger burstiness asks for bigger compensation. The further extensive simulations indicate the limitation of the scheme. When the traffic presents strong bursty feature and at the same time the delay requirement is too strict, the performance of the random compensation service is not good.

3) BOUND OF DELAY-VIOLATION PROBABILITY
We are interested in evaluating the scheme performance under the specific configuration. The difference is that whereas in Fig. 6 the delay constraint of Poisson traffic flow is fixed while the delay constraints of MMOO traffic flow are variable, in this simulation, MMOO traffic flow has fixed on strict delay constraint (20 time slots) while Poisson traffic flow has the variable delay-QoS demand (20 to 160 time slots). The burstiness of MMOO is setting \( c^2 = 1.17 \). Delay-violation probability threshold \( \epsilon = \{10^{-3}, 10^{-4}, 10^{-5}\} \). It can be found from Fig. 7 that compensation bandwidth varies within a narrow range (roughly from 3.6-3.8), also the compensation probabilities keep within the range of 0.005-0.065 roughly. The large-range variation in delay-QoS requirement does not result in strong variation in compensation bandwidth and probability, which reflects the feature of Poisson traffic. The effectiveness of the proposed slicing scheme is again proven, since the solved slicing factor and bandwidth could satisfy the slice reliability.

\[ P( c^2 = 1.5) \]
\[ P( c^2 = 2.0) \]
\[ P( c^2 = 2.5) \]
\[ C( c^2 = 1.5) \]
\[ C( c^2 = 2.0) \]
\[ C( c^2 = 2.5) \]

FIGURE 6. The relationships among burstiness of MMOO, delay constraints of MMOO flow, compensation probability, and compensation bandwidth in NSI 1 (\( \epsilon \) for Poisson flow is 10^{-3}, \( \epsilon \) for MMOO flow is 10^{-3}, \( c^2 = 1.17 \), \( \lambda = R_{av} = 10 \) packets/slot). The delay constraint of Poisson in NSI 2 is 20 time slots.

4) COMBINATION TYPES OF TRAFFIC FLOWS
Customized services may be required in actual 5G network services, so we are curious about the slice configuration gain in our compensation scheme. The final simulations comprise two cases. Case 1: two MMOO flows with heterogeneous delay-QoS constraints share a compensation bandwidth, the simulation results are shown in Fig. 8. Case 2: two Poisson flows with heterogeneous delay-QoS constraints share a compensation bandwidth, the results are shown in Fig. 9. In the two cases, the delay constraint of the traffic flow on NSI 1 is varying, and takes values from
10 to 80 time slots. Delay-QoS constraint of the traffic flow on NSI 2 is fixed to 20 time slots. The basic service rate of each network slice instance is 10 packets/slot. The derived compensation service rate (bandwidth) for two MMOO flows is up to 15 packets/slot and that for two Poisson flows is only 0.79 packets/slot when delay constraint is 10 time slots. Combinations of traffic among network slice instances are heuristic. As Fig. 9 depicts, the quite small compensation bandwidth is enough to support two Poisson flows, so network resource fragmentation could be used for bandwidth compensation. Comparing the compensation bandwidth $C$ in Fig. 5 with Fig. 8, the compensation service rate of bursty traffic can be reduced when NSI 1 and NSI 2 respectively transmit bursty traffic and more stable traffic. Meanwhile, the independence of performances between two NSIs can be guaranteed.

VI. CONCLUSION

In this paper, the proposed random compensation scheme introduced queuing model to simplify the slicing scheme. Although the presented scheme is simple, it could be useful as a starting point to understand practical application of random compensation scheme and develop the support of heterogeneous traffic transmission. Under the statistical delay-QoS constraints, the random compensation service process could serve the traffic without sending pilot frequently and save bandwidth resources. Due to the random compensation bandwidth was shared among slices and the bandwidth was small, the network resource fragments could be used to adjust the resources of the slices dynamically, thereby improving the resource utilization. For the bandwidth allocation at 5G base stations, the deduced slice adjustment factors could guide practical solutions. Especially, the scheme is quite suitable to be implemented in the communication scenarios containing symbiotic traffic flows, like video call or any video and audio dual-track business. Besides, the performance evaluation section showed the proposed random compensation scheme could support the 5G businesses with diversified characteristics. The main limitations of the proposed scheme are that when the traffic becomes very bursty or the delay requirement is too strict, the random compensation service will consume too much bandwidth resources, which means the performance of the scheme will deteriorate. In this situation, the random compensation factor and bandwidth are too large, which is not a typical use case of the scheme. Our solution expected to use small network resource fragments as the bandwidth of compensation service. The future work will introduce Martingale theory to study the random slicing scheme and obtain tighter delay bounds. Martingale theory has great advantages in the analysis of stochastic process, it could help us find tighter bound of delay although it may incur the difficulties in mathematical handling.

REFERENCES

[1] System Architecture for the 5G System, document TS 23.501, 3GPP, Dec. 2018.

[2] G. Sun, Z. T. Gebrekidan, G. O. Boateng, D. Ayepah-Mensah, and W. Jiang, “Dynamic reservation and deep reinforcement learning based autonomous resource slicing for virtualized radio access networks,” IEEE Access, vol. 7, pp. 45758–45772, 2019.

[3] G. Sun, K. Xiong, G. O. Boateng, D. Ayepah-Mensah, G. Liu, and W. Jiang, “Autonomous resource provisioning and resource customization for mixed traffics in virtualized radio access network,” IEEE Syst. J., vol. 13, no. 3, pp. 2454–2465, Sep. 2019.

[4] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, “EMBB-URLLC resource slicing: A risk-sensitive approach,” IEEE Commun. Lett., vol. 23, no. 4, pp. 740–743, Apr. 2019.

[5] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, “5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view,” IEEE Access, vol. 6, pp. 55765–55779, 2018.

[6] R. Kassab, O. Simeone, P. Popovski, and T. Islam, “Non-orthogonal multiplexing of ultra-reliable and broadband services in fog-radio architectures,” IEEE Access, vol. 7, pp. 13035–13049, 2019.
