Delay Analysis for URLLC in 5G Based on Stochastic Network Calculus

***Abstract*—The fifth generation (5G) wireless networks are upcoming to our life. The higher performance requirements are raised to satisfy the needs in modern communication. Ultra-reliable low latency communications (URLLC) is one of the most important scenarios in 5G. URLLC with strict latency and reliability requirements is widely used in some delay-sensitive applications such as self-driving. As the 3GPP claimed, the URLLC is amenable to 99.999% transmission correctness and within 1ms delay bound. How to meet the requirements of reliability and latency is still an open issue. Some academic studies and companies proposed various methods to design URLLC standard, but little effort has been made on applying a theoretical method to analyze the delay bound. Stochastic network calculus is an elegant way to obtain the delay bound based on traffic models and service guarantees. In this paper, we take the character of 5G architecture into account and use the stochastic network calculus to analyze the delay in URLLC. Some factors which can influence on the delay are obtained. Optimizing these factors to reduce the delay will provide valuable guidelines for the early design of URLLC architecture. Finally, numerical results are presented to verify the correctness of the delay analysis.**

***Index Terms*—5G, URLLC, Stochastic Network Calculus, Delay Analysis**

I. Introduction

The 5G era is getting closer to us. 5G communication technology appeared for the first time with the 2018 Pyeongchang Winter Olympics in South Korea. It helps audiences watch the live broadcast continuously and smoothly. According to International Telecommunication Union announced the 5G standard timetable, 5G will start commercially in 2020 [1]. 5G wireless networks are designed to support diverse and complicated scenarios. The third generation partnership project (3GPP) classify these different scenarios into three big categories: enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable low-latency communications (URLLC) [2].

URLLC is widely used in self-driving, mission critical application and some delay sensitive systems. It has stringent requirements in terms of delay and reliability in the 5G New Radio (NR) systems. The key requirements of URLLC as claimed by the 3GPP are to ensure the latency of user plane data less than 1ms for downlink and uplink, meanwhile to keep very high packet reception reliability about 99.999 percent. [3] The stringent delay requirement needs new 5G NR technology to bridge the gap. Though the existing LTE networks can reach the reliability target, but the cost is some dozens of milliseconds time delay. That is far away from the criteria of URLLC. So the delay becomes the chock point and it needs to be solved. Many academies and companies have proposed some engineering solutions to minimize the delay. Such as the HARQ retransmission or grant-free technology. However, how to analyze the generation of time-delay from a theoretical perspective and propose a group of tactics to reduce the delay effectively is an important research subject.

Stochastic network calculus (SNC) theory is very good at delay performance analysis. The SNC is a continuous development method to analyze network traffic characteristic and evaluate performance [4]. Different from queuing theory, the SNC permits some packets violate the desired performance. This feature can better take advantage of statistical multiplexing gains [5]. To deal with random service and statistical guarantee, the SNC theory comes into being with a large number of stochastic processes and network traffic models. Under a suitable traffic model and a chosen server model, the SNC theory can process service guarantee analysis of communication network like delay and backlog. So we capitalize on the SNC method to analyze the delay of the 5G URLLC transmission in this paper. We use stochastic arrival curve to describe the process that user equipment (UE) data sends to gNodeB (gNB) side. According to the 5G network topology architecture, we can deduce the rest stages of data transmission from gNB to cloud server. Every stage of stochastic arrival curve characterizes the delay property, therefore the whole delay of URLLC system is comprised by delays which generated from UE to cloud server. Our main contributions of this paper can be summarized as follows:

1) We build a tandem model to simulate 5G network architecture. In this model, we can analog the data transmission in uplink or downlink from UE to cloud server. We use stochastic service process and concatenation property to analysis the latency.

2) Our analysis results represent which parameters are the key factors affecting the delay. By adjusting the key factors, we give a group of tactics to reduced the delay effectively.

3) Delay analysis and tactics for reducing latency have valuable theoretical guidance for the design of URLLC deployment. In order to meet stringent delay requirements, it provides guidelines for how to allocate resources.

The rest of this paper is organized as follows. Section II summarizes related work of URLLC technology and stochastic network calculus. We present a tandem network model to describe URLLC in Section III. In particular, we illustrate the architecture of this system and analyze the causes of the delay in this section. In section IV, we introduce the experimental environment and analyze the relationship between latency and

main factors. We conclude this paper in Section V. Some theoretical proofs are given in appendix.

II. Related Work

Because the standard of URLLC has not been worked out, many researchers have put forward different solutions for the design of URLLC.

Dozens of researches are focus on how to design and implement URLLC to meet the performance requirements. A design without intervention in the baseband/PHY layer for URLLC is to use interface diversity and integrate multiple communication interfaces. Jimmy and his colleagues propose an analysis framework that combines traditional reliability models with technology-specific latency probability distributions [6]. In this way, they can estimate the performance in terms of latency and reliability in such an integrated communication system. To guarantee a low end-to-end delay with low jitter over combined internet and wireless interfaces, the article [7] presents a new multiple-input multiple-output(MIMO) networked round trip time (RTT) skew control algorithm. This RRT skew controller's advantage is that the controller solves the data flow split problem at the controlling node, since the inner loop control signals are the downlink data rates that fully define the data flow split. Jaya Rao and Sophie Vrzic have propose an approach to adopt packet duplication (PD) method to satisfy the latency and reliability requirements [8]. PD technology generates multiple instances and sends them simultaneously in multiple unrelated channels. The receiver selects the best packets according to the channel condition to achieve better transmission reliability. This PD technique can provide a cost-effective solution without increasing the complexity in the radio access network (RAN).

In terms of resource allocation and energy efficiency, there are also some researches on URLLC. How frequency resource are allocated to send a user data in URLLC scenario? That is an interesting study which plunged by Anand A and De Veciana G [9]. Based on the 5G standard technology Orthogonal Frequency Division Multiple Access (OFDMA), they build a One Shot Transmission model which does not allow retransmission. Adopting queuing theory analysis, they find out a result that a small bandwidth over a longer duration is better than a large swath of bandwidth for short duration in One Shot Transmission system. Green energy saving is getting more and more attention. The article [10] provides a coordinated on-off switching scheme across a set of adjacent gNBs. The gNBs share a sleep schedule among themselves, where gNBs with lower offered traffic and fewer connected UEs select longer OFF durations. This is more energy-efficient with on-off mode than traditional mode on the premise of guaranteeing the time delay.

Because URLLC has strict requirements for delay and reliability, it is very meaningful to evaluate the performance of URLLC. Joachim et al. provide an achievable latency bound evaluation in their article [11]. They compare the worst case RAN transmission latencies for different 5G URLLC configurations. The configuration contains FDD, TDD, frequency numerologies and usage of slots. According the analysis, a frequency with higher numerology can be used to reduce the latency. An article derived from HUAWEI propose a grant-free mode uplink transmission mechanism [12]. Grant-free transmission is a transmission without scheduling request and dynamic grant. This scheme is poised to meet the reliability requirement of URLLC in uplink transmission. By simulating different numbers of active UEs random arriving, the reliability can be improved after adopting the grant-free mode with increasing retransmission.

In order to satisfy the key requirements including latency and reliability, many state-of-the-art solutions have been discussed. These technologies contains fast HARQ retransmission, MIMO, beam forming, diversity interfaces, D2D communication, Ultra Density Network and so on [13] [14] [15] [16]. Some of these technologies can be employed alone to promote the performance, and others need to be combined together to achieve better results. They all mentioned the design of frame structures. That because low latency and high reliability are contradictory. This requires more flexible frequency and time division.

Stochastic network calculus is a very practical tool, and has a good practical effect in performance analysis and theoretical boundary calculation. At present, the most representative academic research teams are Yuming Jiang, Fidler M, Li and Lei from Beijing University of Posts and Telecommunications (BUPT) and Xin Chen from Beijing Information Science & Technology University (BISTU). As shown in Table.I.

TABLE I
Team that study SNC

Team	Research Types	Research field
Jiang	Theory&Application	Delay Evaluation&Wireless
Fidler M	Theory	Service Curve&Estimation
Li&Lei	Application	Wireless communication
Chen	Application	LTE&Femtocells

Yuming Jiang's research field is relatively extensive, including theory and application [17] [19] [18] [20]. The research direction of Fidler M is more theoretical [21] [22] [23] [24]. BUPT and BISTU mostly adopt SNC to analyze performance criteria and quality of service in real networks. The research field of BUPT is mainly in wireless network communication [25] [26] [27] [28]. And the application background of BISTU is concentrated on the resource allocation and LTE networks [29] [30] [31] [32].

In these teams, there are often cooperation. The document [33] is an article of cooperation between Fidler M and Yuming Jiang, which mainly applies SNC theory to analyze the delay boundary of multi-server systems. Jiang also collaborate with BUPT team to evaluate the delay performance in wireless-powered communication system [25].

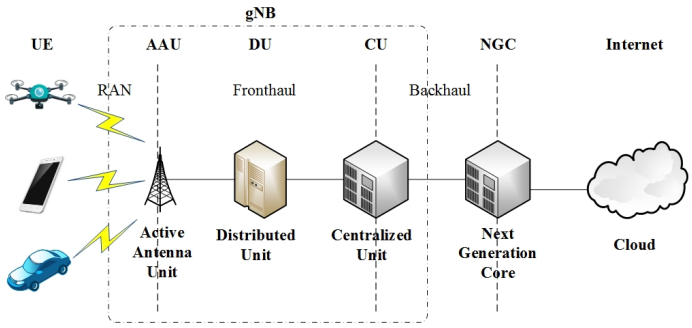

Fig. 1. 5G network architecture.

III. System Model

A. URLLC Network Architecture

We consider URLLC network as a concatenate system from UE to Cloud. We only consider the 5G standalone network situation. The 4G LTE network is composed by UE, RRU (Radio Remote Unit), BBU (Building Base band Unit), the EPC (Evolved Packet Core) which is the LTE's core network, and end by cloud servers.

Different to the 4G LTE, 5G networks are composed by UE, gNB, NGC and Cloud. The gNB contains three parts that are AAU (Active Antenna Unit), DU (Distributed Unit) and CU (Centralized Unit). AAU takes the place of the original RRU and combines some physical layer processing functions of BBU. The BBU function of 4G will be rebuilt into DU and CU in 5G. CU provides the service convergence function in the access side. It focuses on the low real-time capabilities of the protocol stack and adopt a centralized deployment. DU mainly provides data access function to the terminal, including radio frequency and partial signal processing. DU concentrate on the high real-time capabilities of the transport requirements and suit for a distributed deployment method. The NGC (Next Generation Core Network) as Core Network in 5G replace the EPC. 5G NGC is based on SDN/NFV technology and designed to better fit the cloud platform. The architecture is depicted as Fig.1.

B. Delay of URLLC Network

UE devices firstly access to AAU. The AAU is actually a part of base station. This part of the communication belongs to RAN. UE's data will be accepted by AAU, and AAU put forward the data to DU. There are two situations when data arrive at DU. If CU and DU are deployed together, the data can be arrived at CU immediately. Otherwise, the data will be sent to CU from DU. The communication from AAU to CU belongs to fronthaul. The data leave CU and continue upward to the NGC. This part of the communication is called hackhaul. NGC will process the data and it will take some time. Finally, NGC sends the data to Cloud servers. The unidirectional transmission is finished.

So the whole delay or latency in 5G system is contribute by the time processing of RAN, fronthual, backhaul, Core Network and Cloud server. It can be expressed as Formula.1.

$$T_{Total} = T_{RAN} + T_{Fronthaul} + T_{Backhaul} + T_{NGC} + T_{Cloud} \tag{1}$$

where

- T_{RAN} is the time cost by physical layer transmission between UEs and AAU.
- $T_{Fronthaul}$ is the delay between AAU to CU. It is the time taken in gNB.
- $T_{Backhaul}$ is the time taken to communication between gNB to NGC.
- T_{NGC} is the delay taken place in NGC.
- T_{Cloud} is the latency which data transmission between NGC and Cloud server.

To meet the URLLC key requirement, we should do a good job of studying T_{Total}.

In this part, we discuss the latency mainly on the User Plane (UP) rather than Control Plane (C-Plane). That because most of latency is attributed to the UP. UP latency is the communication time between UE and network nodes when transmission and reception of the data at the corresponding IP layer. Whereas, the C-Plane latency is the time spend on radio resource allocating and state switching from idle to active. Compared with UP latency, the C-Plane delay is tiny and can be ignored.

C. Problem Description

The data transport from UEs and go through every nodes to cloud at the end. According to the requirement of URLLC, the reliability and random latency [24] can be described as Formula.2:

$$P\{delay > d\} < \epsilon \tag{2}$$

The delay should be within 1ms, so the $d < 1$, the unit is millisecond (ms). Where ϵ is defined as a very small probability. Meanwhile the block error rate (BLER) must be 99.999%. It means that the value of ϵ must be less than $1 * 10^{-5}$, (i.e. $1 - 1 * 10^{-5} = 0.99999$, It equals to BLER 99.999%). The Formula.2 represents the 5G URLLC network successfully transport data and satisfy the delay and reliability requirements.

D. Stochastic Network Calculus

In SNC theory, the min-plus algebra is applied to analyze queuing system. Let $\mathcal{F}$ denotes the set of non-negative non-decreasing functions and $\bar{\mathcal{F}}$ denotes the set of non-negative non-increasing function. We employ the cumulative process to represent amount of traffic flow. Arrival process, departure process and service process are denoted as $A(t)$, $D(t)$, and $S(t)$ respectively. For any $0 \leqslant s \leqslant t, A(0) = 0, A(s,t) = A(t) - A(s)$, and practical significance of $A(t)$ is the cumulative arrival data at time t. It same to the $D(t)$ and $S(t)$. Some fundamental definitions of curve are well described in literature [4]. We utilize and expand the following in this paper.

***Definition** 1:* **(Stochastic Arrival Curve)**. A flow is said to have a stochastic arrival curve $\alpha \in \mathcal{F}$ with bounding function $f \in \bar{\mathcal{F}}$, denoted by $A(t) \sim< f, \alpha >$, if for all $t \geq 0$ and all $x \geq 0$ there holds

$$P\Big\{ \sup_{0 \leq s \leq t} \{A(s,t) - \alpha(t-s)\} > x \Big\} \leq f(x). \tag{3}$$

where $\alpha(\tau)$ is the stochastic arrival curve, and it denotes the maximum of flow $A(\tau)$. Function $f(x)$ denotes the violation probability. It assumes that the stochastic arrival curve $\alpha(\tau)$ may be exceeded by arrival process $A(\tau)$ in sometimes, but the probability of being exceeded is constrained by the boundary function $f(x)$.

***Definition** 2:* **(Stochastic Service Curve)**. A system S is said to provide a stochastic service curve $\beta \in \mathcal{F}$ with bounding function $g \in \bar{\mathcal{F}}$, denoted by $S \sim< g, \beta >$, if for all $t \geq 0$ there holds

$$P\Big\{ \sup_{0 \leq s \leq t} [A \otimes \beta(s) - D(s)] > x \Big\} \leq g(x). \tag{4}$$

The operator/symbol $\otimes$ represents the cumulative min-plus convolution operation. Which

$$A \otimes \beta(t) = \inf_{0 \leqslant s \leqslant t} \{A(s) + \beta(s,t)\} \tag{5}$$

$\beta(t)$ is the stochastic service curve which means the worst service capability provided by the server. Similar to the stochastic arrival curve, the data that already have been served are probably to be more than the data left. The probability of producing exceeding data can be constrained by the boundary function $g(x)$.

Similarly as in (4), the departure process relates to the arrival and service process and it is described as

$$D(t) \geq \inf_{0 \leqslant s \leqslant t} \{A(s) + S(s,t)\} = A \otimes S(t). \tag{6}$$

where for all $s, t \geqslant 0$ and $s \leqslant t$. That is also the concept of a dynamic server which mentioned in [5]. From the (6), we can better understand the relationship among arrival process, departure process and service process. With these basic processes and curves, we can discuss the definition of the delay boundary.

***Definition** 3:* **(Latency Process)**. Let $A(t)$ and $D(t)$ respectively be the arrival process and departure process. The latency process $L(t)$ at time $t \geq 0$ is defined as

$$L(t) = \inf\{d \geq 0 : A(t) \leq D(t+d)\}. \tag{7}$$

The Formula.7 express that latency $L(t)$ is the least value of d, and the d must meet the condition which the amount of arrival data at time t is less than or equal to the departure data at time $t + d$. It also means that the data do not leave the server immediately. The duration of the data in server is the delay time. In Formula.7, the arrival process $A(t)$ is little than or equal to the departure process $D(t+d)$. It means that the data arrived in server at time t are all leaving from server at time $t + d$. If $A(t)$ is large than or equal to the departure process $D(t+d)$, that represents the data arrived at t moment have not been completed by service during d period of time. So the $A(t)$ little than or equal to $D(t+d)$ situation is utilized to describe the shortest time that server takes for the data to be serviced. That is the latency or delay.

According to the latency process definition, and utilizing stochastic arrival process and stochastic service process, so the stochastic latency bound has been defined at following.

***Theorem** 1:* **(Stochastic Latency Bound)** A system with an input process $A(t)$. $A(t)$ is a stochastic arrival process with stochastic arrival curve $\alpha \in \mathcal{F}$ and bounded by function $f \in \bar{\mathcal{F}}$ (i.e., $A \sim< f, \alpha >$). The system provides to the input a stochastic service process $S(t)$. $S(t)$ is with stochastic service curve $\beta \in \mathcal{F}$ with bounding function $g \in \bar{\mathcal{F}}$ (i.e., $S \sim< g, \beta >$). Then, for all $t \geq 0$ and $x \geq 0$, the Latency $L(t)$ is bounded by

$$P\{L(t) > h(\alpha + x, \beta)\} \leq f \otimes g(x) \tag{8}$$

where function $h(\alpha + x, \beta)$ denotes the maximum horizontal distance between $\alpha + x$ and β, the express $f \otimes g(x)$ represents the cumulative min-plus convolution operation of function f and g.

***Theorem** 2:* **(Concatenation Property)** Considering a flow passes through a network of N server nodes in tandem. If each server nodes $n(= 1, 2, ..., N)$ provides a stochastic service curve $S^n \sim< g^n, \beta^n >$ to its input, then the network guarantees to the flow a stochastic service curve $S \sim< g, \beta >$ with

$$\beta(t) = \beta^1 \otimes \beta^2 \otimes \cdots \otimes \beta^N(t)$$
$$g(x) = g^1 \otimes g^2 \otimes \cdots \otimes g^N(t)$$

E. Model Building

The network character of URLLC can be described as a dynamic server by stochastic processes as introducing in above. The data sent by UE can be described as arrival process $A(t)$ and the service capacity provided by the network server node can be represented by stochastic service process $S(t)$.

As assumption the latency of URLLC in (1), we consider the URLLC network is a tandem system. So the delay of URLLC should fall in the concatenation characterization in SNC.

We consider that a UE's network flow passing through the gNB, CN and Cloud in tandem mode. Each network node $k(= gNB, AAU, DU, CU, CN, Cloud)$ provides a stochastic service curve $S_k \sim< g_k, \beta_k >$ to its flow. We first discuss about the gNB subsystem. The gNB includes AAU, DU and CU, so $S_{AAU}(s,t)$, $S_{DU}(s,t)$ and $S_{CU}(s,t)$ are in series. We

use the same indices to denote the arrival and departure process of the respective systems. Especially the source of data is from UE. So $A_{AAU}(t)$ as arrival process denotes the input data of AAU from UE in gNB subsystem. The arrival process of DU $A_{DU}(t)$ actually equals to the departure process of AAU $D_{AAU}(t)$, where $A_{DU}(t) = D_{AAU}(t)$. By the same token, $A_{CU}(t) = D_{DU}(t)$ similarly for CU server. The departure process of CU is $D_{CU}(t)$. It is also the departure process of the gNB subsystem.

Consider the gNB subsystem is tandem deploy, we assume that the AAU server provides a $S_{AAU}(t)$ capacity to deal with the arrival data. Applying (6), the departure process can be represent as

$$D_{AAU}(t) \geq A_{AAU} \otimes S_{AAU}(t). \tag{9}$$

Similar to the departure process of AAU, we can get the process of DU

$$D_{DU}(t) \geq A_{DU} \otimes S_{DU}(t). \tag{10}$$

Because of $A_{DU}(t) = D_{AAU}(t) = A_{AAU} \otimes S_{AAU}(t)$, we put (9) into (10) to replace $A_{DU}(t)$ by $A_{AAU} \otimes S_{AAU}(t)$ and get

$$D_{DU}(t) \geq (A_{AAU} \otimes S_{AAU}(t)) \otimes S_{DU}(t). \tag{11}$$

By recursive insertion, we can obtain

$$D_{CU}(t) \geq ((A_{AAU} \otimes S_{AAU}) \otimes S_{DU}) \otimes S_{CU}(t). \tag{12}$$

Applying the associativity of min-plus convolution, it holds that

$$\begin{aligned} D_{CU}(t) &\geq ((A_{AAU} \otimes S_{AAU}) \otimes S_{DU}) \otimes S_{CU}(t) \\ &\geq A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t). \end{aligned}$$

From the gNB subsystem perspective to see, $A_{AAU}(t)$ is the first input and $D_{CU}(t)$ is the last output of gNB. So $A_{AAU}(t)$ equal to $A_{gNB}(t)$, and $D_{CU}(t)$ is the departure process of the gNB $D_{gNB}(t)$. Then we can get gNB subsystem

$$D_{gNB} \geq A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t). \tag{13}$$

Assuming first-come first-served order, we use definition.3 and equation (13), and let L_{gNB} denotes the latency process of gNB, there holds

$$\begin{aligned} L_{gNB}(t) &= \inf\{d \geqslant 0 : A_{gNB}(t) - D_{gNB}(t+d) \leqslant 0\} \\ &= \inf\{d \geqslant 0 : A_{AAU}(t) - D_{CU}(t+d) \leqslant 0\} \\ &= \inf\{d \geqslant 0 : A_{AAU}(t) - \\ &\quad A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t+d) \leqslant 0\} \end{aligned} \tag{14}$$

With Theorem.1 and Theorem.2, the delay bound can be analysis by following corollary.

Corollary 1: **(Latency Bound of gNB)** In gNB subsystem, $A_{AAU}(t)$ is a stochastic arrival process with stochastic arrival curve α_{AAU}, i.e. $A \sim< f_{AAU}, \alpha_{AAU} >$. $\alpha_{AAU} \in \mathcal{F}$, $f_{AAU} \in \bar{\mathcal{F}}$. The server nodes in subsystem provide stochastic service process $S_{AAU}(t), S_{DU}(t), S_{CU}(t)$ respectively, i.e. $S_{AAU} \sim< g_{AAU}, \beta_{AAU} >, S_{DU} \sim< g_{DU}, \beta_{DU} >, S_{CU} \sim< g_{CU}, \beta_{CU} >$. And $\beta_{AAU}, \beta_{DU}, \beta_{CU} \in \mathcal{F}$, $g_{AAU}, g_{DU}, g_{CU} \in \bar{\mathcal{F}}$. Then, for all $t \geq 0$ and $x \geq 0$, the Latency of gNB subsystem $L_{gNB}(t)$ is bounded by

$$\begin{aligned} P\{L_{gNB}(t) \geq d\} &= P\{L_{gNB}(t) \geq h(\alpha_{AUU} + x, \beta_{gNB})\} \\ &\leqslant f_{AAU} \otimes g_{gNB}(x) \end{aligned} \tag{15}$$

where service rate $\beta_{gNB}(t) = \beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU}(t)$, and bound function $g_{gNB} = g_{AAU} \otimes g_{DU} \otimes g_{CU}(x)$.

$Proof$: Please see Appendix A.

For mobile edge computing (MEC) deployment, CU can be the end of the transmission. That because the computing resource and storage resource are located at CU. The Corollary.1 is good enough to analysis the delay of communication from UE to CU. However, in order to comprehensively discuss the delay of URLLC system, we need to convert the destination from CU to Cloud. By extending Corollary.1, we can draw the whole URLLC system latency bound.

Corollary 2: **(Latency Bound of URLLC)** In 5G URLLC system, $A_{AAU}(t)$ is a stochastic arrival process with stochastic arrival curve α_{AAU}, i.e. $A \sim< f_{AAU}, \alpha_{gNB} >$. $\alpha_{AAU} \in \mathcal{F}$, $f_{AAU} \in \bar{\mathcal{F}}$. The server nodes in URLLC system provide stochastic service process $S_{gNB}(t)$, $S_{NGC}(t)$ and $S_{Cloud}(t)$ respectively, i.e. $S_{gNB} \sim< g_{gNB}, \beta_{gNB} >$, $S_{NGC} \sim< g_{NGC}, \beta_{NGC} >$, and $S_{Cloud} \sim< g_{Cloud}, \beta_{Cloud} >$. Service rate $\beta_{gNB}, \beta_{NGC}, \beta_{Cloud} \in \mathcal{F}$, $g_{gNB}, g_{NGC}, g_{Cloud} \in \bar{\mathcal{F}}$. Then, for all $t \geq 0$ and $x \geq 0$, the Latency of URLLC system $L_{All}(t)$ is bounded by

$$\begin{aligned} P\{L_{All}(t) \geq d\} &= P\{L_{All}(t) \geq h(\alpha_{AUU} + x, \beta_{All})\} \\ &\leqslant f_{AAU} \otimes g_{All}(x) \end{aligned} \tag{16}$$

where $\beta_{All}(t) = \beta_{gNB} \otimes \beta_{NGC} \otimes \beta_{Cloud}(t)$, and $g_{All} = g_{gNB} \otimes g_{NGC} \otimes g_{Cloud}(x)$.

$Proof$: Please see Appendix B.

We have built a model to represent the latency of 5G MEC (UE to gNB) and URLLC (the full path from UE to Cloud). Next we intend to calculate the delay boundary of the model. With Corollary.2, we know that 4 key variance need to be determined. These are stochastic arrival process α_{AAU}, f_{AAU}, and stochastic service process β_{All}, g_{All}. Especially, we can also decompose β_{All} and g_{All} to obtain more detailed result.

In URLLC scenario, the data is usually fixed unit packet size, the data size sometimes very tiny while applying millimeter-wave technology. We assume that UE data arrive will approximate to a Poisson distribution with mean rate λ. So arrive curve $\alpha_{AAU} = \lambda t$ and bound function will be

$$f_{AAU}(x) = \sum_{k=x+\lambda t}^{\infty} \frac{e^{\lambda t} \cdot (\lambda t)^k}{k!} \tag{17}$$

With MGF of right hand in (17) and Chernoff bound, f_{AAU} can be tighten by

$$f_{AAU}(x) = e^{x-(\lambda t+x)ln\frac{\lambda t+x}{\lambda t}} \tag{18}$$

The proof of this part can be found in [20]. Two variances in stochastic arrival process have been solved. We begin to determine β_{All}, g_{All} for the stochastic service process.

In order to simplify the problem, we generalize service rate of server nodes and assume that each node provides data processing capacity as $\beta(t) = Ct$ with bounding function $g(x) = ae^{-bx}$. According to Corollary.2, we can get

$$g_{All}(x) = g_{AAU} \otimes g_{DU} \otimes g_{CU} \otimes g_{NGC} \otimes g_{Cloud}(x) \quad (19)$$

Therefore,

$$\begin{aligned} g_{All}(x) &= g_{AAU} \otimes g_{DU} \otimes g_{CU} \otimes g_{NGC} \otimes g_{Cloud}(x) \\ &= \inf_{x_1+x_2+x_3+x_4+x_5=x} \sum_{k=1}^{5} a_k e^{-b_k x_k} \end{aligned} \quad (20)$$

Applying with the conclusion in [18], we can hold

$$\begin{aligned} &\inf_{x_1+x_2+x_3+x_4+x_5=x} \sum_{k=1}^{5} a_k e^{-b_k x_k} \\ &= e^{\frac{-x}{w}} \prod_{k=1}^{5} (a_k b_k w)^{\frac{1}{b_k w}} \end{aligned} \quad (21)$$

where $w = \sum_{k=1}^{5} \frac{1}{b_k}$, and service bound functions respectively are $g_{AAU}(x) = a_1 e^{-b_1 x_1}$, $g_{DU}(x) = a_2 e^{-b_2 x_2}$, $g_{CU}(x) = a_3 e^{-b_3 x_3}$, $g_{NGC}(x) = a_4 e^{-b_4 x_4}$, $g_{Cloud}(x) = a_5 e^{-b_5 x_5}$. with all the information we discuss above, and applying the lemma which proved in [18], we can get

$$g_{All}(x) = e^{\frac{-xb}{n+1}}(a(n+1)) \quad (22)$$

We put (18), (22) into (16) and apply the Theorem.3 proved in [34], it can be derived that

$$\begin{aligned} &P\{L(t) > h(\alpha_{AAU} + x, \beta)\} \\ &= P\{L(t) > \frac{x}{C-\lambda}\} \\ &\leq e^{x-(\lambda t+x) ln \frac{\lambda t+x}{\lambda t}} \cdot e^{\frac{-xb}{n+1}}(a(n+1)) \end{aligned} \quad (23)$$

Then we hold the latency bound as (2). Let $d = \frac{x}{C-\lambda}$ and set the right side of (23) equal to ϵ. The ϵ is a small latency bound violation probability. We can obtain a relationship between d and ϵ.

$$d = \frac{1}{C-\lambda} \cdot \frac{n+1}{b} \cdot ln \frac{a(n+1)}{\epsilon} \quad (24)$$

The calculation process can be found in Appendix C.

IV. Numerical Result and Performance Evaluation

In this section, we shall discuss what factors are the main cause of latency in URLLC.

According to the delay requirement of URLLC, we write a simulator to validate our analysis result. We implement the simulator in Python 3.6. It can calculate the delay value in accordance with the input network parameters.

To validate the correctness of analysis, we set the violation probability from 1×10^{-6} to 1×10^{-4}. We compare the analysis with the simulator, and the Figure.2 shows the result. Figure.2 shows a good match between the analytical data and the corresponding simulation result. Especially, the analytical result is also the boundary of the simulation data. The boundary is slightly loose when the value of violation probability is large, while it is more tighter when value is little. This fully illustrates the correctness of the analysis results.

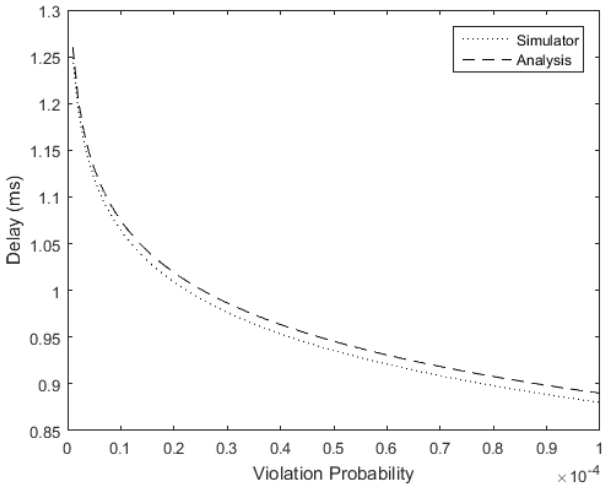

Fig. 2. Comparison between Simulator and Analysis

Although the deployment details of the URLLC standard are not yet released, we can still apply SNC theory for quantitative analysis.

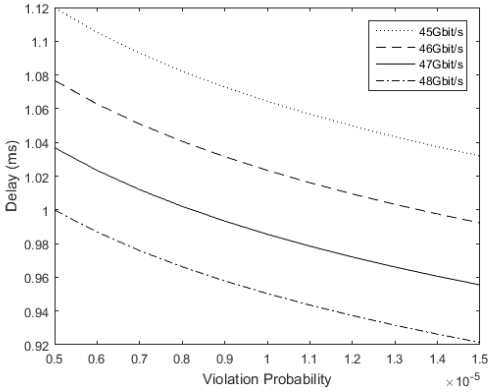

Fig. 3. Service Rate Influence

We assume that the 5G URLLC networks are standalone deployment. The packet arrival rate is constant and arrival process satisfies Poisson distribution. A general URLLC reliability requirement for one transmission of a packet is $1*10^{-5}$ for 32 bytes with a user plane latency of 1ms. So we set the violation probability value around $1*10^{-5}$. More simulation parameters can be found in Table.II.

TABLE II
EVALUATION PARAMETERS

Parameter	Value
Network deployment	Standalone
Traffic mode	Constant transmission, Poisson arrival
Carrier frequency	4 (GHz)
Arrival rate λ	20 (Gbit/s)
Service rate C	40,45,50,55 (Gbit/s)
Service bound a	1
Service bound b	3
Number of tandem servers n	5
Violation probability	$1 * 10^{-5}$
Latency bound	1 (ms)

Taking this boundary probability as the precondition, we simulate the relationship between system latency and service rate by applying the conclusions we have drawn in the previous section. Fig.3 provide the evaluated URLLC delay with different service rate under violation probability $1*10^{-5}$. The violation probability represent the BLER. The value of BLER is from $5 * 10^{-6}$ to $15 * 10^{-6}$. We arrange the value scope to include the demand value $1 * 10^{-5}$ to observe the effect of this value on delay. We adopt 4 service rates in model and all the curves are slow down by BLER value. From this we can conduct that BLER is not the main factor to influence the latency. In order to make the delay less than 1ms, we set service rate from 45Gbit/s to 48Gbit/s based on arrival rate 20Gbit/s. We can procure that delay approximates 1ms when service rate is 47Gbit/s at violation probability $1*10^{-5}$. As the service rate increases, the delay of the system will decrease. When service rate is 48Gbit/s, system latency can approach 1ms with lower violation probability. That means system can guarantee the low latency communication in a stable state.

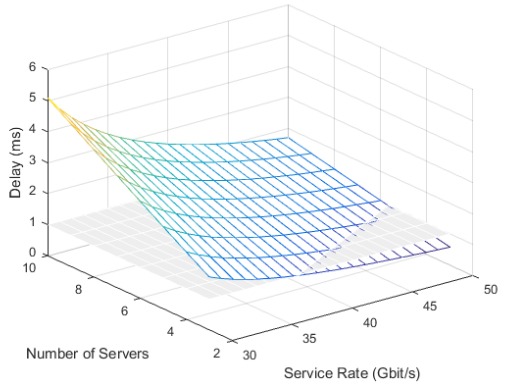

Fig. 4. Number of Servers Influence

Fig.4 presents the relationship among latency, number of servers and service rate. The arrival rate is constant and speed is 20Gbit/s. The violation probability is maintained at $1*10^{-5}$. Based on the above setting, we can derived that the delay is sensitive on number of servers in tandem. From Fig.4, we can see the slope of latency cause by number of server in tandem is larger than the service rate. We draw a delay equals 1ms flat plane to cut the curve surface. The part below the plane is the scope of deployment parameters which satisfying the delay condition. We can also see that in order to ensure low latency of communication, it is necessary to reduce the levels of system deployment as much as possible and increase the service rate of each server layer.

V. Conclusions

In this paper, the architecture of 5G URLLC network researched. According to the architecture characteristics, the URLLC network is modeled as a tandem system which demonstrate the communication from UE to Cloud. Applying stochastic network theory and combining the features of URLLC network, performance analysis has been conducted. We have investigated the relationship among: the delay constraint, the service rate, violation probability and the number of deployment servers levels in URLLC networks. The 3GPP standard is taken into account when we set the simulation parameters. Numerical results are demonstrate that the main factor which can impact on latency is number of server deployment levels. That also means Edge Computing will be the trend in URLLC application deployment. The service rate of the server is also a factor affecting the delay. With the increase of service rate, delay can be reduced. The results derived from evaluation which provide valuable guidelines for the early design of URLLC deployment. With the further development of the researches, we would consider to include handover access in URLLC communication.

VI. Appendix

A. Proof of Corollary.1

$Proof$: Since the latency process Definition.(3) are defined as $L(t) = \inf\{d \geq 0 : A(t) \leq D(t+d)\}$, event $L(t) > d$ implies event $A(t) \leq D(t+d)$. We move $D(t+d)$ from right hand to left hand, and according to (14), the latency bound of gNB can be hold as

$$P\{L_{gNB}(t) > d\} \leqslant P\{A_{AAU}(t) - D_{CU}(t+d) \leq 0\} \quad (25)$$

Then we focus on the $\{A_{AAU}(t) - D_{CU}(t+d)\}$ part. We put right hand of (13) into this part, we can get

$$\begin{aligned} &A_{AAU}(t) - D_{CU}(t+d) \\ =&A_{AAU}(t) - A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t+d) \\ +&A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t+d) - D_{CU}(t+d) \end{aligned} \quad (26)$$

With the Theorem.2, utilizing the concatenation property we can obtain that $S_{AAU} \sim< g_{AAU}, \beta_{AAU} >$, $S_{DU} \sim< g_{DU}, \beta_{DU} >$, $S_{CU} \sim< g_{CU}, \beta_{CU} >$. Stochastic service process convolution operation $(S_{AAU} \otimes S_{DU} \otimes S_{CU})$ means gNB subsystem provides maybe lower than $(\beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU})$ processing capacity, but the violation probability in this case is limited by $g_{AAU} \otimes g_{DU} \otimes g_{DU}$. Through applying (4), we

denote β_{gNB} equals to $(\beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU})$, g_{gNB} equals to $g_{AAU} \otimes g_{DU} \otimes g_{DU}$. hence (26) can hold be

$$\begin{aligned}
&A_{AAU}(t) - D_{CU}(t+d) \\
=&A_{AAU}(t) - A_{AAU} \otimes (\beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU})(t+d) \\
&+ A_{AAU} \otimes (\beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU})(t+d) - D_{CU}(t+d) \\
=&A_{AAU}(t) - A_{AAU} \otimes \beta_{gNB}(t+d) \\
&+ A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d)
\end{aligned} \tag{27}$$

According to (5), we replace $A_{AAU} \otimes \beta_{gNB}(t+d)$ by $\inf\{A_{AAU}(s) + \beta_{gNB}(t+d-s)\}$ in (27). Consequently,

$$\begin{aligned}
&A_{AAU}(t) - D_{CU}(t+d) \\
=&A_{AAU}(t) - A_{AAU} \otimes \beta_{gNB}(t+d) \\
&+ A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d) \\
=&A_{AAU}(t) - \inf_{0 \leqslant s \leqslant t+d} \{A_{AAU}(s) + \beta_{gNB}(t+d-s)\} \\
&+ A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d) \\
\leq&A_{AAU}(t) - A_{AAU}(s) - \beta_{gNB}(t+d-s)\} \\
&+ A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d) \\
\leq&A_{AAU}(s,t) - \alpha_{AAU}(t-s) \\
&+ \alpha_{AAU}(s,t) - \beta_{gNB}(t+d-s) \\
&+ A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d)
\end{aligned} \tag{28}$$

We add α_{AAU} at step 4 in (28) to build stochastic arrival curve. Based on the stochastic arrival curve (3),$A_{AAU}(s,t) - \alpha_{AAU}(s,t)$ is less than or equal to f_{AAU}. Applying stochastic service curve (4), $A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d)$ is less than or equal to g_{gNB}. With Theorum.1, we use $h(\alpha + x, \beta)$ replace the d. where $h(\alpha + x, \beta)$ is the maximum horizontal distance between $\alpha + x$ and β for $x \geq 0$. The $h(\alpha, \beta)$ function implies the condition

$$\lim_{t \to \infty} [\alpha(t) - \beta(t)] \leq 0. \tag{29}$$

we can obtain

$$\begin{aligned}
&P\{L(t) > h(\alpha_{AAU} + x, \beta_{gNB})\} \\
&= P\{\{A_{AAU}(t) - D_{CU}(t + h(\alpha + x, \beta))\} > 0\} \\
&\leqslant \sup_{0 \leqslant s \leqslant t} \{A_{AAU}(s,t) - \alpha_{AAU}(t-s)\} \\
&+ \sup_{0 \leqslant s \leqslant t + h(\alpha_{AAU} + x, \beta_{gNB})} \{A_{AAU} \otimes \beta_{gNB}(s) - D_{CU}(s)\} \\
&\leqslant f_{AAU}(t) + g_{gNB}(x) \\
&\leqslant \inf\{f_{AAU}(t) + g_{gNB}(x-t)\} \\
&\leqslant f_{AAU} \otimes g_{gNB}(x)
\end{aligned}$$

Therefore, Corollary.1 is proved.

B. Proof of Corollary.2

Proof : In the Corollary.1, the gNB subsystem are constituted by AAU, DU and CU. In addition to gNB, the whole 5G URLLC system also include NGC and Cloud. According to the concatenation property which mentioned in Theorem.2, then the network guarantees to the flow a stochastic service curve $S_{All} \sim < g_{All}, \beta_{All} >$ with

$$\beta_{All}(t) = \beta_{gNB} \otimes \beta_{CN} \otimes \beta_{Cloud}(t) \tag{30}$$

where

$$\beta_{gNB}(t) = \beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU}(t) \tag{31}$$

actually

$$\beta_{All}(t) = \beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU} \otimes \beta_{CN} \otimes \beta_{Cloud}(t) \tag{32}$$

and

$$g_{All}(x) = g_{gNB} \otimes g_{CN} \otimes g_{Cloud}(x) \tag{33}$$

where

$$g_{gNB}(x) = g_{AAU} \otimes g_{DU} \otimes g_{CU}(x) \tag{34}$$

actually

$$g_{All}(x) = g_{AAU} \otimes g_{DU} \otimes g_{CU} \otimes g_{CN} \otimes g_{Cloud}(x) \tag{35}$$

Based on latency process Definition.(3), the 5G URLLC system latency process can be defined as

$$L(t) = \inf\{d \geqslant 0 : A_{AAU}(t) \leq D_{Cloud}(t+d)\} \tag{36}$$

Latency bound of 5G URLLC is defined as

$$P\{L(t) \geq d\} = P\{A_{AAU}(t) - D_{Cloud}(t+d) \leq 0\} \tag{37}$$

We also focus on $A_{AAU}(t) - D_{Cloud}(t+d)$ part. where $D_{Cloud} \geqslant A_{AAU} \otimes S_{AAU} \otimes S_{DU} \otimes S_{CU} \otimes S_{NGC} \otimes S_{Cloud}$. Then we have

$$\begin{aligned}
&A_{AAU}(t) - D_{Cloud}(t+d) \\
=&A_{AAU}(t) - A_{AAU} \otimes S_{AAU} \otimes S_{DU} \otimes S_{CU} \\
&\otimes S_{NGC} \otimes S_{Cloud}(t+d) \\
&+ A_{AAU} \otimes S_{AAU} \otimes S_{DU} \otimes S_{CU} \\
&\otimes S_{NGC} \otimes S_{Cloud} - D_{Cloud}(t+d) \\
=&A_{AAU}(t) - A_{AAU} \otimes S_{gNB} \otimes S_{NGC} \otimes S_{Cloud}(t+d) \\
&+ A_{AAU} \otimes S_{gNB} \otimes S_{NGC} \otimes S_{Cloud} - D_{Cloud}(t+d) \\
\leq&A_{AAU}(t) - A_{AAU}(s) \\
&- \beta_{gNB} \otimes \beta_{NGC} \otimes \beta_{Cloud}(t+d-s)) \\
&+ A_{AAU} \otimes S_{gNB} \otimes S_{NGC} \otimes S_{Cloud} - D_{Cloud}(t+d) \\
\leq&A_{AAU}(s,t) - \alpha_{AAU}(t-s) \\
&+ \alpha_{AAU}(t-s) - \beta_{all}(t+d-s) \\
&+ A_{AAU} \otimes \beta_{all}(t+d) - D_{Cloud}(t+d)
\end{aligned}$$

With stochastic arrival curve (3), $A_{AAU}(s,t) - \alpha(t-s)$ is bounded by $f_{AAU}(x)$. According to stochastic service curve (4), $A_{AAU} \otimes \beta_{All}(t+d) - D_{Cloud}(t+d)$ is limited by g_{All}. For long-term running, if $t \to \infty$, $\alpha_{AAU}(t-s) - \beta_{All}(t+d-s)$ approximate to zero because of $\alpha_{AAU}, \beta_{All} \in \mathcal{F}$. Finally, with Theorem.1, the delay of the URLLC system can be bounded by this

$$P\{L(x) > h(\alpha_{AAU} + x, \beta_{All})\} < f_{AUU} \otimes g_{All}(x) \tag{38}$$

Therefore, Corollary.2 is proved.

C. Calculation of Delay

We first set right side of (23) equals to ϵ, and we logarithm on both sides then hold

$$\begin{aligned} e^{x-(\lambda t+x)ln\frac{\lambda t+x}{\lambda t}} \cdot e^{\frac{-xb}{n+1}}(a(n+1)) &= \epsilon \\ e^{x-(\lambda t+x)ln\frac{\lambda t+x}{\lambda t}} \cdot e^{\frac{-xb}{n+1}} &= \frac{\epsilon}{a(n+1)} \end{aligned} \tag{39}$$

for a long-term running situation, $t \to \infty$, then

$$\lim_{t\to\infty}(\lambda t+x)ln\frac{\lambda t+x}{\lambda t} = x \tag{40}$$

then the (39) equal to

$$\begin{aligned} e^{\frac{-xb}{n+1}} &= \frac{\epsilon}{(a(n+1))} \\ \frac{-xb}{n+1} &= ln\frac{\epsilon}{(a(n+1))} \\ x &= \frac{n+1}{b}\cdot ln\frac{a(n+1)}{\epsilon} \end{aligned} \tag{41}$$

we put $x = d(C-\lambda)$ into (41) and get

$$d = \frac{1}{C-\lambda}\cdot\frac{n+1}{b}\cdot ln\frac{a(n+1)}{\epsilon} \tag{42}$$

Therefore d is solved.

ACKNOWLEDGMENT

This work was supported by these programs:
National Natural Science Foundation of China (Nos.61370065,61502040),
National Key Technology Research and Development Program of the Ministry of Science and Technology of China (No.2015BAK12B03-03),
Beijing Municipal Program for Excellent Teacher Promotion (No.PXM2017_014224.000028).

REFERENCES

[1] ITU-R M.2083-0, IMT Vision - Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond, Sept 2015.
[2] 3GPP TR 38.913, Study on Scenarios and Requirements for Next Generation Access Technologies, June 2017.
[3] Soldani D, Guo Y J, Barani B, et al, 5G for Ultra-Reliable Low-Latency Communications, IEEE Network, 2018, vol.32, no.2, pp.6-7.
[4] Jiang Y, Liu Y. Stochastic Network Calculus, Springer London, 2009.
[5] M. Fidler and A. Rizk, A Guide to the Stochastic Network Calculus, in IEEE Communications Surveys & Tutorials, vol. 17, no. 1, pp. 92-105, Firstquarter 2015. pp.92-105.
[6] J. J. Nielsen, R. Liu and P. Popovski, Ultra-Reliable Low Latency Communication Using Interface Diversity, IEEE Transactions on Communications, vol.66, no.3, pp.1322-1334, March 2018
[7] Delgado R A, Lau K, Middleton R H, et al. Networked Delay Control for 5G Wireless Machine-Type Communications Using Multiconnectivity, IEEE Transactions on Control Systems Technology, vol.99, pp.1-16, 2018.
[8] Rao J, Vrzic S, Packet Duplication for URLLC in 5G: Architectural Enhancements and Performance Analysis, IEEE Network, vol.32, no.2, pp.32-40, 2018.
[9] Anand A, De Veciana G, Resource Allocation and HARQ Optimization for URLLC Traffic in 5G Wireless Networks, 2018.
[10] Mukherjee A, Energy Efficiency and Delay in 5G Ultra-Reliable Low-Latency Communications System Architectures, IEEE Network, 2018, 32(2):55-61.
[11] Sachs J, Wikstrom G, Dudda T, et al, 5G Radio Network Design for Ultra-Reliable Low-Latency Communication, IEEE Network, 2018, vol.32, no.2, pp.24-31.
[12] Wang C, Chen Y, Wu Y, et al. Performance Evaluation of Grant-Free Transmission for Uplink URLLC Services, IEEE, Vehicular Technology Conference: Vtc2017-Spring. IEEE, 2017:1-6.
[13] Pocovi G, Shariatmadari H, Berardinelli G, et al. Achieving Ultra-Reliable Low-Latency Communications: Challenges and Envisioned System Enhancements, IEEE Network, vol.32, no.2, pp.8-15, 2018.
[14] Popovski P, Nielsen J J, Stefanovic C, et al. Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks, IEEE Network, vol.32, no.2, pp.16-23, 2018.
[15] Ji H, Park S, Yeo J, et al. Introduction to Ultra Reliable and Low Latency Communications in 5G, 2017.
[16] Ji H, Park S, Yeo J, et al. Ultra Reliable and Low Latency Communications in 5G Downlink: Physical Layer Aspects, 2018.
[17] Z. Li, Y. Jiang, Y. Gao, P. Li, L. Sang and D. Yang, Delay and Delay-Constrained Throughput Performance of a Wireless-Powered Communication System, IEEE Access, vol.5, pp.21620-21631, 2017.
[18] Sun F, Li L, Jiang Y. Impact of duty cycle on end-to-end performance in a Wireless Sensor Network, Wireless Communications and NETWORKING Conference IEEE, pp.1906-1911, 2015.
[19] Jiang Y. Network calculus and queueing theory: two sides of one coin: invited paper, International ICST Conference on PERFORMANCE Evaluation Methodologies & TOOLS, 2014.
[20] Wu K, Jiang Y, Li J. On the model transform in stochastic network calculus, International Workshop on Quality of Service IEEE, pp.1-9, 2010.
[21] S. Akin and M. Fidler, A Method for Cross-Layer Analysis of Transmit Buffer Delays in Message Index Domain, IEEE Transactions on Vehicular Technology, vol.67, no.3, pp.2698-2712, March 2018.
[22] R. Lbben and M. Fidler, Service Curve Estimation-Based Characterization and Evaluation of Closed-Loop Flow Control, IEEE Transactions on Network and Service Management, vol.14, no.1, pp.161-175, March 2017.
[23] R. Lbben and M. Fidler, Estimation method for the delay performance of closed-loop flow control with application to TCP, IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications, San Francisco, CA, pp.1-9, 2016.
[24] R. Lbben, M. Fidler and J. Liebeherr, Stochastic Bandwidth Estimation in Networks With Random Service, IEEE/ACM Transactions on Networking, vol.22, no.2, pp.484-497, April 2014.
[25] Z. Li, Y. Jiang, Y. Gao, P. Li, L. Sang and D. Yang, Delay and Delay-Constrained Throughput Performance of a Wireless-Powered Communication System, IEEE Access, vol.5, pp.21620-21631, 2017.
[26] Z. Li, Y. Gao, B. A. Salihu, et al, Network Calculus Delay Bounds in Multi-Server Queueing Networks with Stochastic Arrivals and Stochastic Services, IEEE Global Communications Conference, pp.1-7, 2015.
[27] K. Zheng, F. Liu, L. Lei, C. Lin and Y. Jiang, Stochastic Performance Analysis of a Wireless Finite-State Markov Channel, in IEEE Transactions on Wireless Communications, vol.12, no.2, pp.782-793, February 2013.
[28] Y. Li, L. Lei, Z. Zhong and S. Lin, Performance analysis for high-speed railway communication network using stochastic network calculus, 5th IET International Conference on Wireless, Mobile and Multimedia Networks (ICWMMN 2013), Beijing, pp.100-105, 2013.
[29] Chen, Xin, Y. Si, and X. Xiang. Delay-bounded resource allocation for femtocells exploiting the statistical multiplexing gain, Kluwer Academic Publishers, 2015.
[30] X. Chen, Y. Si and X. Xiang, Delay-bounded priority-driven resource allocation for two-tier macrocell/femtocell downlink, The 2014 5th International Conference on Game Theory for Networks, Beijing, pp.1-6, 2014.
[31] Tong Xu, Xin Chen, Xudong Xiang and Jianxiong Wan, Delay analysis for cognitive radio networks with parallel Markov modulated on-off channels, ISWPC 2012 proceedings, Dalian, 2012, pp. 1-5.
[32] Lei Zhang, Xin Chen, Xudong Xiang and Jianxiong Wan, A stochastic network calculus approach for the end-to-end delay analysis of LTE networks, 2011 International Conference on Selected Topics in Mobile and Wireless Networking (iCOST), Shanghai, pp.30-35, 2011.
[33] M. Fidler, B. Walker and Y. Jiang, Non-Asymptotic Delay Bounds for Multi-Server Systems with Synchronization Constraints, in IEEE Transactions on Parallel and Distributed Systems, vol.29, no.7, pp.1545-1559, July 1 2018.
[34] Beck M. Towards the analysis of transient phases with stochastic network calculus, Telecommunications Network Strategy and Planning Symposium IEEE, pp.164-169, 2016.