

Performance evaluation of URLLC in 5G based on Stochastic Network Calculus

Shengcheng Ma · Xin Chen · Zhuo Li ·
Ying Chen

Received: date / Accepted: date

Abstract 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 delay-sensitive application such as self-driving. As the 3GPP claims, 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. Few efforts have been made to 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. A tandem model describing the communication in the 5G network is built, and parameters which have an influence on the delay are analyzed. Numerical results are presented to verify the correctness of the delay analysis. We investigate how to optimize the parameters to reduce the delay, which would provide valuable guidelines for the design of URLLC architecture.

Supported by program National Natural Science Foundation of China (Nos.61872044,61502040). Beijing Municipal Program for Excellent Teacher Promotion (No.PXM2017_014224.000028). Beijing Municipal Program for Top Talent Cultivation (CIT&TCD201804055). Qinxin Talent Program of Beijing Information Science and Technology University.

F. Shengcheng Ma
Beijing Information Science and Technology University, Beijing, China
Tel.: +86-18510412330
E-mail: mashengcheng@163.com

S. Xin Chen
Beijing Information Science and Technology University, Beijing, China

T. Zhuo Li
Beijing Information Science and Technology University, Beijing, China E-mail: l-izhuo@bistu.edu.cn

F. Ying Chen
Beijing Information Science and Technology University, Beijing, China

Keywords 5G · URLLC · Stochastic Network Calculus · Delay Analysis

1 Introduction

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 the 5G standard schedule announced by the International Telecommunication Union (ITU), 5G will begin commercialization in 2020 [1]. 5G wireless networks are designed to support diverse and complicated scenarios. The third generation partnership project (3GPP) classifies 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, meanwhile to keep reliability not less than 99.999 percent [3]. The stringent delay requirement needs new 5G NR technology to bridge the gap. Although existing LTE networks can achieve reliability goals, the cost is some dozens of milliseconds of delay. That is far away from the criteria of URLLC. So the delay becomes the bottleneck and this problem 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 delay from a theoretical perspective and propose a strategy to effectively reduce the delay is an important research subject.

Stochastic network calculus (SNC) theory is an effective tool on delay performance analysis. SNC is a continuous development method to analyze network traffic characteristic and evaluate performance[4]. Different from queuing theory, SNC permits some packets to violate the desired performance. This feature can better take advantage of statistical multiplexing gains[5]. To deal with random service and statistical guarantee, SNC theory comes into being with a large number of stochastic processes and network traffic models. SNC theory can process service guarantee analysis of communication networks, such as 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 generated from UE to the 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 analyze the data transmission from UE to cloud server. We use stochastic service process and concatenation property to analysis the latency.

2) For delay analysis, stochastic latency bound for URLLC network is obtained. To analyze the reliability of the network, we provide a stochastic error process which can express the block error rate. We also investigate the impacts of block error rate on latency, and a delay boundary analysis method considering error rate is proposed.

3) Based on our analysis results, we can find out what parameters are the key factors affecting the delay. Delay analysis and strategy 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.

2 Related Work

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

Researches 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 analytical 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 round trip time (RTT) skew control controller with multiple-input multiple-output(MIMO). This controller's advantage is that it solves the data flow split problem at the controlling node. Jaya Rao and Sophie Vrzić have proposed 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 packet according to the channel condition in order 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 the frequency resource will be allocated to the

user to send data in URLLC scenario. That is an interesting study which is 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. 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. If the gNBs have lower traffic and fewer connected UEs, they will be set to the OFF mode. This On-Off mode is more energy-efficient 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 latencies for different configurations in 5G RAN transmission. The configuration contains FDD, TDD, frequency numerologies and usage of slots. According to the analysis, a frequency with higher numerology can be used to reduce the latency. An article derived from HUAWEI company proposes a grant-free mode uplink transmission mechanism[12]. Grant-free transmission grants dynamically without scheduling request. This mode is poised to meet the reliability requirement of URLLC in uplink transmission. By simulating random arriving from different numbers of active UEs, 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, some state-of-the-art solutions have been discussed in [13][14][15][16]. These technologies contain fast HARQ retransmission, MIMO, beam forming, diversity interfaces, D2D communication, Ultra Density Network and so on. Some of these technologies can be employed alone to promote 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 it has a good practical effect in performance analysis and theoretical boundary calculation. Li and Jiang [17] analyzed the throughput performance on the wireless-powered communication system. They considered the delay as a constrained condition. The stochastic traffic arrival was adopted to derive the cumulative data transmission capacity. M. Fidler and his partner [18] used stochastic service process to analyze the end-to-end delay performance of TCP. The estimation method of closed-loop flow was implemented by random service process in this paper. They considered both backlog and delay by using stochastic network calculus. In article [19], the performance of wireless finite-state Markov channel was analyzed. The delay boundary was derived based on the moment generating functions (MGF). Xin Chen's team was focused on LTE network and researched the resource allocation to guarantee the delay performance [20]. The delay was constrained by the difference value between stochastic arrival

curve and stochastic service curve. The article [21] is an article of cooperation between Fidler M and Jiang. It mainly applies SNC theory to analyze the delay boundary of multi-server systems.

3 System Model

3.1 URLLC Network Architecture

Generally speaking, 5G network includes 5G Standalone networking (SA) and Non-standalone networking (NSA). The SA mode is the establishment of whole new 5G network, while the NSA mode is the combination of 5G and 4G LTE. According to the implementation technology of core network, the NSA mode can continue to be divided. In order to simplify the system model, we only discuss the case of 5G standalone network.

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

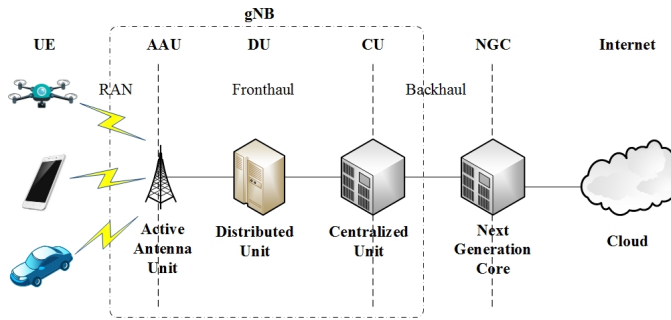


Fig. 1: 5G network architecture.

3.2 Delay of URLLC Network

UE devices firstly access to AAU. The AAU is actually a part of the 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 arrive 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 backhaul. 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 contributed by the time processing of RAN, fronthaul, backhaul, Core Network, and Cloud server. It can be expressed as (1).

$$T_{Total} = T_{RAN} + T_{Fronthaul} + T_{Backhaul} + T_{NGC} + T_{Cloud} \quad (1)$$

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 the Control Plane (C-Plane). That because most of the 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. The results of comparison between UP latency and C-Plane have been discussed in article [10]. To simplify the analysis, we mainly consider the delay generated by the User Plane.

3.3 Problem Description

The data is transferred from UE, through each node, and finally to the cloud. According to the requirement of URLLC, the random latency [22] can be described as (2):

$$P\{\text{delay} > d\} < \epsilon \quad (2)$$

The delay should be within 1ms, so the $d < 1$, the unit is millisecond (ms). Where ϵ is defined as a very small probability. The (2) represents the 5G URLLC network successfully transport data and satisfy the delay requirements.

3.4 Basics of 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 $0 \leq s \leq t$, $A(0) = 0$, $A(s, t) = A(t) - A(s)$, and practical significance of $A(t)$ is the cumulative arrival data at time t . It is the same for $D(t)$ and $S(t)$. Some fundamental definitions of curve are well described in the 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

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

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

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

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

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

$\beta(t)$ 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 processed are probably more than the data departed. 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

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

Where for all $s, t \geq 0$ and $s \leq t$. That is also the concept of a dynamic server which mentioned in [3]. 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

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

The (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 (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.

Lemma 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., $A \sim \langle f, \alpha \rangle$). 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., $S \sim \langle g, \beta \rangle$). Then, for all $t \geq 0$ and $x \geq 0$, the Latency $L(t)$ is bounded by*

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

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 .

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

$$\begin{aligned} \beta(t) &= \beta^1 \otimes \beta^2 \otimes \dots \otimes \beta^N(t) \\ g(x) &= g^1 \otimes g^2 \otimes \dots \otimes g^N(t) \end{aligned} \quad (9)$$

3.5 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 represented by the arrival process $A(t)$. The service capacity provided by the network server node can be depicted as the stochastic service process $S(t)$.

As assumption the latency of URLLC in Formula.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 $S_k \sim \langle g_k, \beta_k \rangle$ 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.

3.5.1 Delay of gNB

Considering that the gNB subsystem is tandem deployed, 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

$$D_{AAU}(t) \geq A_{AAU} \otimes S_{AAU}(t). \quad (10)$$

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

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

Because of $A_{DU}(t) = D_{AAU}(t) = A_{AAU} \otimes S_{AAU}(t)$, we put (10) into (11) to replace $A_{DU}(t)$ by $A_{AAU} \otimes S_{AAU}(t)$ and get

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

By recursive insertion, we can obtain

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

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

$$\begin{aligned} D_{CU}(t) &\geq ((A_{AAU} \otimes S_{AAU}) \otimes S_{DU}) \otimes S_{CU}(t) \\ &\geq A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t). \end{aligned} \quad (14)$$

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

$$D_{gNB} \geq A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t). \quad (15)$$

Assuming first-come first-served order, we use Definition.3 and equation (15), and let L_{gNB} denotes the latency process of gNB, there holds

$$\begin{aligned}
L_{gNB}(t) &= \inf\{d \geq 0 : A_{gNB}(t) - D_{gNB}(t+d) \leq 0\} \\
&= \inf\{d \geq 0 : A_{AAU}(t) - D_{CU}(t+d) \leq 0\} \\
&= \inf\{d \geq 0 : A_{AAU}(t) \\
&\quad - A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t+d) \leq 0\}
\end{aligned} \tag{16}$$

With Lemma.1 and Lemma.2, the delay bound can be analysis by following theorem.

Theorem 1 (Latency Bound of gNB) In gNB subsystem, $A_{AAU}(t)$ is a stochastic arrival process with stochastic arrival curve α_{AAU} , i.e. $A \sim \langle f_{AAU}, \alpha_{AAU} \rangle$. $\alpha_{AAU} \in \mathcal{F}$, $f_{AAU} \in \bar{\mathcal{F}}$. The server nodes in subsystem provide stochastic service process $S_{AAU}(t), S_{DU}(t), S_{CU}(t)$ respectively, i.e. $S_{AAU} \sim \langle g_{AAU}, \beta_{AAU} \rangle$, $S_{DU} \sim \langle g_{DU}, \beta_{DU} \rangle$, $S_{CU} \sim \langle g_{CU}, \beta_{CU} \rangle$. 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

$$\begin{aligned}
P\{L_{gNB}(t) \geq d\} &= P\{L_{gNB}(t) \geq h(\alpha_{AAU} + x, \beta_{gNB})\} \\
&\leq f_{AAU} \otimes g_{gNB}(x)
\end{aligned} \tag{17}$$

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 : 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 (16), the latency bound of gNB can be hold as

$$P\{L_{gNB}(t) > d\} \leq P\{A_{AAU}(t) - D_{CU}(t+d) \leq 0\} \tag{18}$$

Then we focus on the $\{A_{AAU}(t) - D_{CU}(t+d)\}$ part. We put right hand of (15) into this part, we can get

$$\begin{aligned}
&A_{AAU}(t) - D_{CU}(t+d) \\
&= A_{AAU}(t) - A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t+d) \\
&\quad + A_{AAU} \otimes (S_{AAU} \otimes S_{DU} \otimes S_{CU})(t+d) - D_{CU}(t+d)
\end{aligned} \tag{19}$$

With the Lemma.2, utilizing the concatenation property we can obtain that $S_{AAU} \sim \langle g_{AAU}, \beta_{AAU} \rangle$, $S_{DU} \sim \langle g_{DU}, \beta_{DU} \rangle$, $S_{CU} \sim \langle g_{CU}, \beta_{CU} \rangle$. 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 $g_{AAU} \otimes g_{DU} \otimes g_{CU}$. hence (19) can hold be

$$\begin{aligned}
& A_{AAU}(t) - D_{CU}(t+d) \\
&= A_{AAU}(t) - A_{AAU} \otimes (\beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU})(t+d) \\
&\quad + A_{AAU} \otimes (\beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU})(t+d) - D_{CU}(t+d) \\
&= A_{AAU}(t) - A_{AAU} \otimes \beta_{gNB}(t+d) \\
&\quad + A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d)
\end{aligned} \tag{20}$$

According to (5), we replace $A_{AAU} \otimes \beta_{gNB}(t+d)$ by $\inf\{A_{AAU}(s) + \beta_{gNB}(t+d-s)\}$ in (20). Consequently,

$$\begin{aligned}
& A_{AAU}(t) - D_{CU}(t+d) \\
&= A_{AAU}(t) - A_{AAU} \otimes \beta_{gNB}(t+d) \\
&\quad + A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d) \\
&= A_{AAU}(t) - \inf_{0 \leq s \leq t+d} \{A_{AAU}(s) + \beta_{gNB}(t+d-s)\} \\
&\quad + A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d) \\
&\leq A_{AAU}(t) - A_{AAU}(s) - \beta_{gNB}(t+d-s) \\
&\quad + A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d) \\
&\leq A_{AAU}(s, t) - \alpha_{AAU}(t-s) \\
&\quad + \alpha_{AAU}(s, t) - \beta_{gNB}(t+d-s) \\
&\quad + A_{AAU} \otimes \beta_{gNB}(t+d) - D_{CU}(t+d)
\end{aligned} \tag{21}$$

We add α_{AAU} at step 4 in (21) 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 Theorem.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

$$\lim_{t \rightarrow \infty} [\alpha(t) - \beta(t)] \leq 0. \tag{22}$$

we can obtain

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

Therefore, Theorem.1 is proved.

3.5.2 Delay of URLLC

For mobile edge computing (MEC) deployment, CU can be the end of the transmission. That is because the computing resource and storage resource are located at CU. The Theorem.1 is suitable for 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 Theorem.1, we can draw the whole URLLC system latency bound.

Theorem 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. $A \sim \langle f_{AAU}, \alpha_{gNB} \rangle$. $\alpha_{AAU} \in \mathcal{F}$, $f_{AAU} \in \bar{\mathcal{F}}$. The server nodes in URLLC system provide stochastic service process $S_{gNB}(t)$, $S_{NGC}(t)$ and $S_{Cloud}(t)$ respectively, i.e. $S_{gNB} \sim \langle g_{gNB}, \beta_{gNB} \rangle$, $S_{NGC} \sim \langle g_{NGC}, \beta_{NGC} \rangle$, and $S_{Cloud} \sim \langle g_{Cloud}, \beta_{Cloud} \rangle$. 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*

$$\begin{aligned} P\{L_{All}(t) \geq d\} &= P\{L_{All}(t) \geq h(\alpha_{AAU} + x, \beta_{All})\} \\ &\leq f_{AAU} \otimes g_{All}(x) \end{aligned} \quad (23)$$

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 : In the Theorem.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 Lemma.2, then the network guarantees to the flow a stochastic service curve $S_{All} \sim \langle g_{All}, \beta_{All} \rangle$ with

$$\beta_{All}(t) = \beta_{gNB} \otimes \beta_{CN} \otimes \beta_{Cloud}(t) \quad (24)$$

where

$$\beta_{gNB}(t) = \beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU}(t) \quad (25)$$

actually

$$\beta_{All}(t) = \beta_{AAU} \otimes \beta_{DU} \otimes \beta_{CU} \otimes \beta_{CN} \otimes \beta_{Cloud}(t) \quad (26)$$

and

$$g_{All}(x) = g_{gNB} \otimes g_{CN} \otimes g_{Cloud}(x) \quad (27)$$

where

$$g_{gNB}(x) = g_{AAU} \otimes g_{DU} \otimes g_{CU}(x) \quad (28)$$

actually

$$g_{All}(x) = g_{AAU} \otimes g_{DU} \otimes g_{CU} \otimes g_{CN} \otimes g_{Cloud}(x) \quad (29)$$

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

$$L(t) = \inf\{d \geq 0 : A_{AAU}(t) \leq D_{Cloud}(t + d)\} \quad (30)$$

Latency bound of 5G URLLC is defined as

$$P\{L(t) \geq d\} = P\{A_{AAU}(t) - D_{Cloud}(t+d) \leq 0\} \quad (31)$$

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

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

With stochastic arrival curve (3), $A_{AAU}(s, t) - \alpha_{AAU}(t-s)$ 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 $t \rightarrow \infty$, $\alpha_{AAU}(t-s) - \beta_{All}(t+d-s)$ approximate to zero because of $\alpha_{AAU}, \beta_{All} \in \mathcal{F}$. Finally, with Lemma.1, the delay of the URLLC system can be bounded by this

$$P\{L(x) > h(\alpha_{AAU} + x, \beta_{All})\} < f_{AAU} \otimes g_{All}(x) \quad (32)$$

Therefore, Theorem.2 is proved.

3.5.3 Reliability of URLLC

In evaluating the reliability of the system, the stochastic network calculus theory still has no perfect solution. Some researchers are studying the system loss, which reflects the reliability of the system from another angle. In term of loss analysis in stochastic network calculus, article [23] and [24] are both representative research on loss process. A. Gulyas and J. Biro propose a Workload Loss Ratio Estimation to express the loss in the network flow. An effective w-arrival curve Z^φ is put forward in this paper. They use $Z^\varphi \odot S^{\varphi_s}$ to present the loss generated in service process. However, whether it is accurate to describe the actual bandwidth by means of effective w-service curve remains to be discussed. In paper [24], the authors give a stochastic loss bound by MGF. The loss is considered in arrival process which is the beginning of the transmission. The arrival process is split by a loss process $L(t)$ and a lossless process

$R(t)$. That means $A(t) = L(t) + R(t)$, and the real flow to the server is $R(t)$. And the main contribution for this paper is to obtain a tighter stochastic loss bound of $L(t)$. By contrast, the loss of traffic from the server node is more reasonable than from the arrival process. In URLLC communication, we will mainly focus on the errors in processing and checking on the server. Referring to the article [25], a novel stochastic curve which work on server node will be proposed.

In order to study the reliability of URLLC, BLER is an important criteria to be discussed. We propose a new curve to describe the block error rate in URLLC network. We denote the stochastic error process $I(\tau)$ as the block error rate occurred in the server.

Definition 4 (Stochastic Error Curve). A system is said to have a stochastic error curve $\gamma \in \mathcal{F}$ with bounding function $h \in \bar{\mathcal{F}}$, denoted by $I \sim_{<} h, \gamma >$, if for all $t \geq 0$ there holds

$$P\left\{\sup_{0 \leq s \leq t} [I(s, t) - \gamma(t - s)] > x\right\} \leq h(x). \quad (33)$$

The (33) means that the system server handles the received blocks and probably checks out the erroneous data. The process $I(s, t)$ represents the statistical value of block error rate generated from time s to t . The curve γ is the ceiling for the error process I . The error process I is usually less than the error curve γ . However, we assume that the error curve $\gamma(\tau)$ may be exceeded by error process $I(\tau)$ in sometimes, but the probability of being exceeded is constrained by the boundary function $h(x)$.

In order to express the ultra reliability of the URLLC communication, we can set the γ value as the BLER. If the value of γ is small, it means the running state of system is stable and errors rarely occurs. In stochastic service curve (4), $A \otimes \beta$ represent the flow with lossless situation. When we take into account the BLER in network system, γ is the upper bound of the block error rate. And $1 - \gamma$ is the correct block rate. The $\hat{\beta}$ denotes the service rate without errors. So $A \otimes (\hat{\beta} * (1 - \gamma))$ denotes the service capacity after considering BLER. We can simplify $A \otimes (\hat{\beta} * (1 - \gamma))$ as $A \otimes \beta$, where $\beta = \hat{\beta} * (1 - \gamma)$.

MGF is widely used in stochastic network calculus to compute the performance bound for network system. MFG has a beneficial characteristics that it is convenient to get the sum of stochastic processes by product form. We compute the stochastic error bound and analyze the BLER with MGF. The MGF of stochastic error process is defined as (34) with any $\theta \leq 0$

$$M_I(\theta, t - s) = E[e^{\theta I(s, t)}] \quad (34)$$

And with the Chernoff's theory, the stochastic error bound will be

$$P\{I(s, t) \geq x\} \leq e^{-\theta x} E[e^{\theta I(s, t)}] \quad (35)$$

3.5.4 Calculation of Delay and Reliability

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 Theorem.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

$$f_{AAU}(x) = \sum_{k=x+\lambda t}^{\infty} \frac{e^{\lambda t} \cdot (\lambda t)^k}{k!} \quad (36)$$

With MGF of right hand in (36) and Chernoff bound, f_{AAU} can be tighten by

$$f_{AAU}(x) = e^{x - (\lambda t + x) \ln \frac{\lambda t + x}{\lambda t}} \quad (37)$$

The proof of this part can be found in [26]. 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 Theorem.2, we can get

$$g_{All}(x) = g_{AAU} \otimes g_{DU} \otimes g_{CU} \otimes g_{NGC} \otimes g_{Cloud}(x) \quad (38)$$

Therefore,

$$\begin{aligned} g_{All}(x) &= g_{AAU} \otimes g_{DU} \otimes g_{CU} \otimes g_{NGC} \otimes g_{Cloud}(x) \\ &= \inf_{x_1+x_2+x_3+x_4+x_5=x} \sum_{k=1}^5 a_k e^{-b_k x_k} \end{aligned} \quad (39)$$

Applying with the conclusion in [27], we can hold

$$\inf_{x_1+x_2+x_3+x_4+x_5=x} \sum_{k=1}^5 a_k e^{-b_k x_k} = e^{\frac{-x}{w}} \prod_{k=1}^5 (a_k b_k w)^{\frac{1}{b_k w}} \quad (40)$$

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 [27], we can get

$$g_{All}(x) = e^{\frac{-x}{n+1}} (a(n+1)) \quad (41)$$

We put (37), (41) into (23) and apply the Theorem.3 proved in [28], it can be derived that

$$\begin{aligned} & P\{L(t) > h(\alpha_{AAU} + x, \beta)\} \\ &= P\{L(t) > \frac{x}{C - \lambda}\} \\ &\leq e^{x - (\lambda t + x) \ln \frac{\lambda t + x}{\lambda t}} \cdot e^{\frac{-xb}{n+1}} (a(n+1)) \end{aligned} \quad (42)$$

Then we hold the latency bound as (2). Let $d = \frac{x}{C - \lambda}$ and set the right side of (42) equal to ϵ . The ϵ is a small latency bound violation probability. We can obtain a relationship between d and ϵ .

$$d = \frac{1}{C - \lambda} \cdot \frac{n+1}{b} \cdot \ln \frac{a(n+1)}{\epsilon} \quad (43)$$

We first set right side of (42) equals to ϵ , and we logarithm on both sides then hold

$$\begin{aligned} & e^{x - (\lambda t + x) \ln \frac{\lambda t + x}{\lambda t}} \cdot e^{\frac{-xb}{n+1}} (a(n+1)) = \epsilon \\ & e^{x - (\lambda t + x) \ln \frac{\lambda t + x}{\lambda t}} \cdot e^{\frac{-xb}{n+1}} = \frac{\epsilon}{a(n+1)} \end{aligned} \quad (44)$$

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

$$\lim_{t \rightarrow \infty} (\lambda t + x) \ln \frac{\lambda t + x}{\lambda t} = x \quad (45)$$

then the (44) equal to

$$\begin{aligned} & e^{\frac{-xb}{n+1}} = \frac{\epsilon}{a(n+1)} \\ & \frac{-xb}{n+1} = \ln \frac{\epsilon}{a(n+1)} \\ & x = \frac{n+1}{b} \cdot \ln \frac{a(n+1)}{\epsilon} \end{aligned} \quad (46)$$

we put $x = d(C - \lambda)$ into (46) and get

$$d = \frac{1}{C - \lambda} \cdot \frac{n+1}{b} \cdot \ln \frac{a(n+1)}{\epsilon} \quad (47)$$

Therefore d is solved.

Considering the reliability of URLLC, the service rate of system β should be impacted by BLER. We study on the right hand of (35) to obtain the bound of stochastic error process. By reference article [5], we adopt the envelope model and get

$$E[e^{\theta I(t)}] \leq e^{\theta(\gamma(t) + \sigma)} \quad (48)$$

Table 1: evaluation parameters

Parameter	Value
Network deployment	Standalone
Traffic mode	Constant transmission, Poisson arrival
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
BLER	$1 * 10^{-5}$
Latency bound	1 (ms)

Then we put (48) into right hand of (35), and we hold

$$\begin{aligned}
P\{I(s, t) \geq x\} &\leq e^{-\theta x} \cdot e^{\theta(\gamma(t-s)+\sigma)} \\
&\leq e^{-\theta x + \theta(\gamma(t-s)+\sigma)} \\
&\leq e^{\theta(\gamma(t-s)+\sigma-x)}
\end{aligned} \tag{49}$$

For the above delay analysis, the service rate needs to be adjusted to accommodate the occurrence of error rate.

$$\hat{C} = C(1 - \gamma) \tag{50}$$

where C is the service capacity without loss and error, and the \hat{C} is the adjusted service rate. Then we can get a stochastic delay bound with block error rate.

$$\begin{aligned}
&P\{L(t) > h(\alpha_{AAU} + x, \beta)\} \\
&= P\{L(t) > \frac{x}{C(1 - \gamma) - \lambda}\} \\
&= P\{L(t) > \frac{x}{\hat{C} - \lambda}\} \leq e^{\frac{-xb}{n+1}}(a(n+1))
\end{aligned} \tag{51}$$

4 Numerical Results and Performance Evaluation

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

Although the deployment details of the URLLC standard are not yet released, we can still apply SNC theory for quantitative analysis. 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 block error rate value around $1 * 10^{-5}$. More simulation parameters can be found in Table.1.

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.2 provide the evaluated URLLC

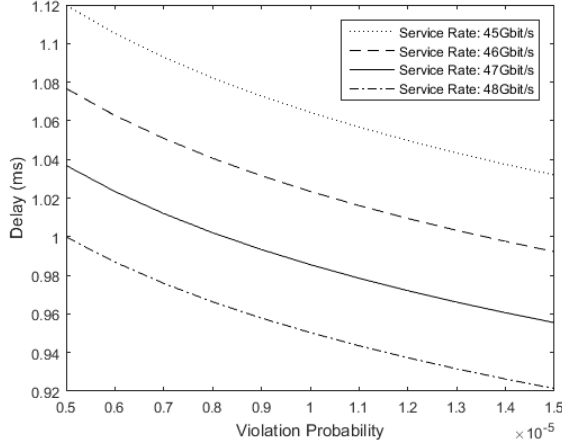


Fig. 2: Service Rate Influence

delay with different service rate under violation probability 1×10^{-5} . The value of violation probability is from 5×10^{-6} to 15×10^{-6} . We arrange the value scope to observe the effect of this value on delay. We adopt 4 service rates in model and all the curves are slow down by violation probability value. From this, we can conduct that the violation probability 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.

For Reliability of URLLC, we adopt SNC theory to study the stochastic bound of block error rate. In Fig.(3), we set the parameter γ in three different values which is near to the 1×10^{-5} . This parameter is set according to the BLER requirement of URLLC in 3GPP[2]. We can get that the block error rate will get down with the violate probability increasing. For the upper bound parameter $\gamma = 0.00001$, the BLER is well limited to the 1×10^{-5} . It also means that the 99.999% transmission correctness.

Based on the reliability demand of URLLC, we study the influence of block error rate on system delay. From Fig.4, we can see that BLER has little effect on the delay. The delay varies very slightly when BLER value ranges from 1×10^{-5} to 1×10^{-3} . A reason is that ultra reliability requires BLER not to exceed 1×10^{-5} . It is difficult for such a small value to have a large impact on service rate. Thus the reliability of the system is guaranteed.

Fig.5 illustrates the relationship between system delay and service rate and arrival rate. As shown in the figure, system delay raises with the increase of arrival rate. When the arrival rate is 20Gbit/s, the system delay exceeds

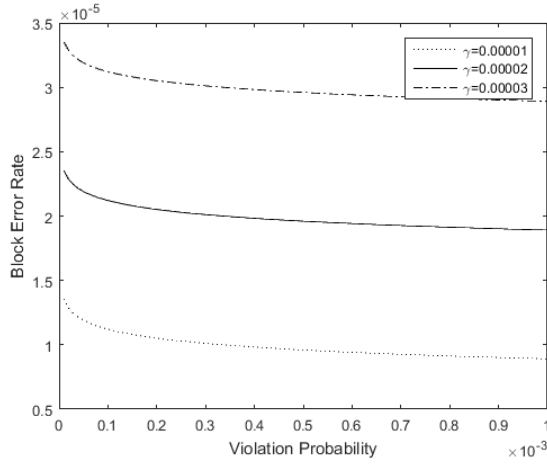
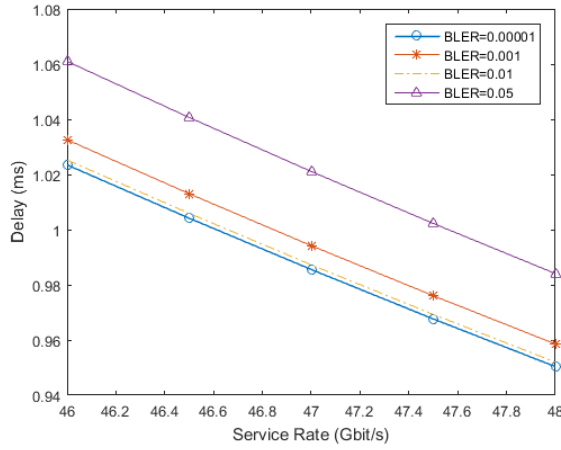
Fig. 3: BLER Bound with parameter γ 

Fig. 4: Delay with Different BLER

1ms which is the requirement of URLLC. With the increase of service rate, the system delay can be reduced to less than 1ms. But when the arrival rate continues to increase, the delay will eventually exceed the demand. Due to limited resources, service rates cannot increase infinitely. Therefore, we should do a good job in task scheduling and resource allocation.

Fig.6 presents the relationship among latency, number of server levels and service rate. The arrival rate is constant and the 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 tandem servers. From Fig.6, we can see the slope of latency caused by number of tandem servers is larger

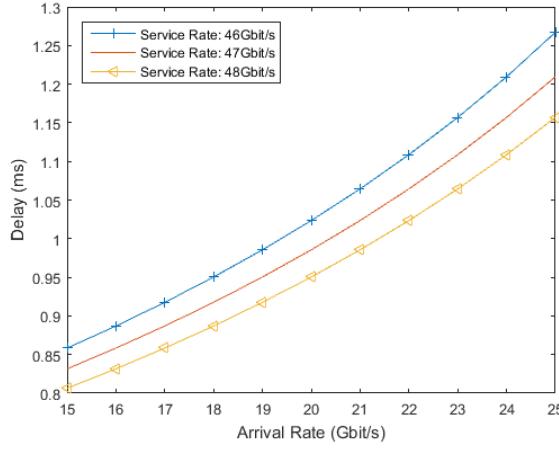


Fig. 5: Delay with Arrival Rate

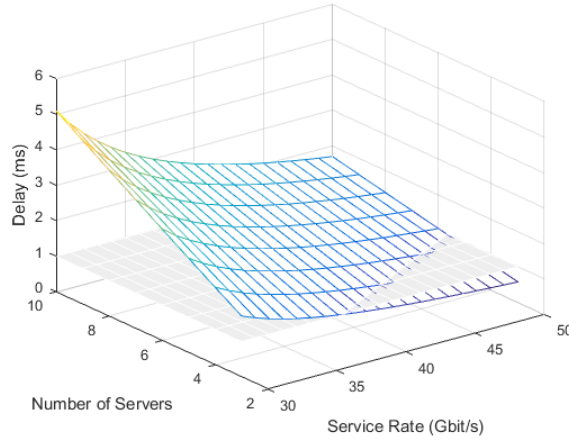


Fig. 6: Number of Tandem Servers Influence

than the service rate. We draw a delay equals 1ms flat plane to cut the curved surface. The part below the plane is the scope of deployment parameters which satisfy the delay condition. In order to ensure low latency of communication, it is necessary to reduce the number of tandem servers deployment as much as possible and increase the service rate of each server layer.

5 Conclusions

In this paper, the architecture of the 5G URLLC network is researched. According to the architecture characteristics, the URLLC network is modeled

as a tandem system which describes 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 between delay constraint, service rate, violation probability, block error rate, and the number of deployed servers in URLLC network. The 3GPP standard is taken into account when we set the simulation parameters. Numerical results verify that the main factor which can impact latency is the number of servers deployed in tandem. 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, the delay can be reduced. The results derived from evaluation provide valuable guidelines for the early design of URLLC deployment. For our future work, we would consider to include handover access in URLLC communication.

References

1. ITU-R M.2083-0, IMT Vision - Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond (2015)
2. 3GPP TR 38.913, Study on Scenarios and Requirements for Next Generation Access Technologies (2017)
3. Soldani D, Guo Y J, Barani B, et al., 5G for Ultra-Reliable Low-Latency Communications, *IEEE Network*, **32**(2), pp.6–7 (2018)
4. Jiang Y, Liu Y, *Stochastic Network Calculus*, Springer, London (2009)
5. M. Fidler and A. Rizk, A Guide to the Stochastic Network Calculus, *IEEE Communications Surveys & Tutorials*, **17**(1), pp.92–105 (92–105)
6. J. J. Nielsen, R. Liu and P. Popovski, Ultra-Reliable Low Latency Communication Using Interface Diversity, *IEEE Transactions on Communications* **66**(3), pp.1322–1334 (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* **99**, pp.1–16 (2018)
8. Rao J, Vrzic S, Packet Duplication for URLLC in 5G: Architectural Enhancements and Performance Analysis, *IEEE Network* **32**(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* **32**(2), pp.55–61 (2018)
11. Sachs J, Wikstrom G, Dudda T, et al., 5G Radio Network Design for Ultra-Reliable Low-Latency Communication, *IEEE Network*, **32**(2), pp.24–31 (2018)
12. Wang C, Chen Y, Wu Y, et al., Performance Evaluation of Grant-Free Transmission for Uplink URLLC Services, *IEEE Vehicular Technology Conference 2017 Vtc2017-Spring*, pp.1–6 (2017)
13. Pocovi G, Shariatmadari H, Berardinelli G, et al., Achieving Ultra-Reliable Low-Latency Communications: Challenges and Envisioned System Enhancements, *IEEE Network*, **32**(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*, **32**(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*, **5**, pp.21620–21631 (2017)

18. 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, pp.1–9 (2016)
19. K. Zheng, F. Liu, L. Lei, C. Lin and Y. Jiang, Stochastic Performance Analysis of a Wireless Finite-State Markov Channel, IEEE Transactions on Wireless Communications **12**(2), pp.782–793 (2013)
20. Chen, Xin, Y. Si, and X. Xiang, Delay-bounded resource allocation for femtocells exploiting the statistical multiplexing gain, **71**, pp.3217–3236 (2015).
21. M. Fidler, B. Walker and Y. Jiang, Non-Asymptotic Delay Bounds for Multi-Server Systems with Synchronization Constraints, IEEE Transactions on Parallel and Distributed Systems, **29**(7), pp.1545–1559 (2018)
22. R. Lbben, M. Fidler and J. Liebeherr, Stochastic Bandwidth Estimation in Networks With Random Service, IEEE/ACM Transactions on Networking, **22**(2), pp.484–497 (2014)
23. Gulys, Andrs, Br, Jzsef, A stochastic extension of network calculus for workload loss examinations, Communications Letters IEEE, **10**(5), pp.399–401 (2006)
24. Deng Y, Lin C, An Extended Stochastic Loss Bound with Moment Generating Function, International Conference on Communications & Mobile Computing IEEE (2010)
25. Sami Ayyorgun, et al., A Composable Service Model With Loss and a Scheduling Algorithm, Infocom (2004)
26. Wu K, Jiang Y, Li J, On the model transform in stochastic network calculus, In: International Workshop on Quality of Service IEEE, pp.1–9 (2010)
27. 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)
28. Beck M, Towards the analysis of transient phases with stochastic network calculus, Telecommunications Network Strategy and Planning Symposium IEEE, pp. 164–169 (2016)