# End-to-End Latency Evaluation of the Sat5G Network Based on Stochastic Network Calculus

**Huaifeng Shi[1], Chengsheng Pan[1, *], Li Yang[2], Debin Wei[2] and Yunqing Shi[3]**

**Abstract:** Simultaneous use of heterogeneous radio access technologies to increase the performance of real-time, reliability and capacity is an inherent feature of satellite-5G integrated network (Sat5G). However, there is still a lack of theoretical characterization of whether the network can satisfy the end-to-end transmission performance for latency-sensitive service. To this end, we build a tandem model considering the connection relationship between the various components in Sat5G network architecture, and give an end-to-end latency calculation function based on this model. By introducing stochastic network calculus, we derive the relationship between the end-to-end latency bound and the violation probability considering the traffic characteristics of multimedia. Numerical results demonstrate the impact of different burst states and different service rates on this relationship, which means the higher the burst of arrival traffic and the higher the average rate of arrival traffic, the greater the probability of end-to-end latency violation. The results will provide valuable guidelines for the traffic control and cache management in Sat5G network.

**Keywords:** 5G, satellite network, stochastic network calculus, latency evaluation.

## 1 Introduction

Benefiting from the rapid development of satellite link technologies and the on-board processing, there will be more than one hundred high-throughput satellite systems using geostationary earth orbit (GEO) and mega-constellations of low earth orbit (LEO) satellites by 2020-2025 [Giambene, Kota and Pillai (2018)]. These evolving satellite systems are expected to provide radio access networks (RANs), which will be integrated into 5G systems along with other wireless technologies [3GPP TR22.822 (2018)]. Seamless handover between heterogeneous radio access technologies will be an inherit characteristic of 5G, while different radio access technologies will be used to improve reliability, availability, capacity and security [Liu, Zeng, Shi et al. (2019)]. Moreover, due to the pressing demand for communication and connection quality, the amount of data produced

[1] School of Automation, Nanjing University of Science and Technology, Nanjing, 210094, China.

[2] School of Information Engineering, Dalian University, Dalian, 116622, China.

[3] Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA.

[*] Corresponding Author: Chengsheng Pan. Email: pancs@sohu.com.

Received: 14 April 2020; Accepted: 31 May 2020.

CMC. doi:10.32604/cmc.2020.011005

each day is growing exponentially [Yu, Liang, He et al. (2018)]. In the June 2018 edition of The Ericsson Mobile report, it is predicted that there will be about 3.5 billion mobile connections by 2023, growing at a rate of 30% a year [Mattisson (2018)].

Faced with the transmission requirements of massive data in the satellite-5G integrated network (Sat5G), how to meet the QoS requirements of services, especially latency-sensitive services, has become an important challenge. However, the traditional TCP/IP protocol is designed to apply "best effort" to optimize the throughput of network traffic and ensure its communication reliability, ignoring the real-time requirements of end-to-end transmission for latency-sensitive services [Liu, Peng, Wang et al. (2019)]. Therefore, it is an important research topic to analyse and evaluate the end-to-end transmission latency theoretically and select the appropriate link to meet the service latency requirements.

Stochastic network calculus (SNC) is a theoretical tool for analysing latency performance [Jiang and Liu (2009); Wang, Zhang, Zhou et al. (2019)]. Unlike queuing theory, the former allows certain traffic to violate expected performance which makes full use of statistical multiplexing gains [Fidler and Rizk (2015)]. For the sake of dealing with QoS guarantee, SNC provides many stochastic process and network traffic models such as latency and backlog of communication network [Ma, Chen, Li et al. (2019)]. Therefore, SNC method is used to analyse Sat5G transmission latency in this paper.

Our main contributions can be summarized as follows:

1) A tandem model is built to simulate the architecture of Sat5G network. By this model, the process of traffic transmission from user equipment (UE) to data network (DN) is analysed, and the end-to-end latency calculation function is derived.
2) By considering the traffic characteristics of multimedia, a self-similar traffic model is built as the stochastic network calculus arrival curve. On this basis, the relationship between the end-to-end latency bound and violation probability is derived.
3) By setting different parameter values, numerical results demonstrate the impact of different burst states and different service rates on this relationship. The results also provide guidelines for deployment of the Sat5G network to meet stringent latency requirements.

The rest of this paper is organized as follows. Section 2 expounds the related work of stochastic network calculus and its application, especially in terms of end-to-end latency evaluation. A tandem network model simulating Sat5G is presented in Section 3. In Section 4, we give a description of the end-to-end latency evaluation problem and establish a latency bound model of Sat5G. Then, we give a relational expression between the end-to-end latency requirement and its violation probability. In Section 5, we introduce the experiment and analyse the relationship of latency and its main influencing factors. Conclusion is in Section 6.

## 2 Related work

Network calculus emerged in the 1990s as a deterministic theory of packet data network service quality analysis [Ma, Chen, Li et al. (2019)]. The arrival flow of the system is simulated by the upper envelope function. The minimum service guarantee provided by a system (such as a router or scheduler) is represented by a so-called service curve

[Burchard, Liebeherr and Patek (2012)]. Based on these concepts, the data departures of a network element can be calculated by convolving the arrival curve with the system service curve [Chang (2000)]. This form of convolution is important because it provides a general framework for the analysis of the entire network. The systems along the network path can easily be connected through the convolution of their service curves to produce network service curves that specify the end-to-end availability of services.

However, with the continuous evolution of communication systems, especially considering the many random factors in wireless network communication, deterministic network calculus is greatly limited in its use [Wang, Di, Jiang et al. (2017)]. Therefore, many scholars worldwide are actively conducting research into stochastic network calculus theory, constantly improving the relevant theories of the theory and applying it to the performance analysis of actual communication systems.

One of the methods of performance analysis is to build a traffic model which is able to accurately describe the traffic characteristics [Cheng, Zhuang and Ling (2007)]. Markov chain model and associated queueing analysis are widely and deeply studied as tools to evaluate the performance of multimedia applications. However, a large number of network traffic measurement studies show that network traffic is self-similar or long-range dependent (LRD) [Izabella, Zhong and Cees (2020)], which cannot be captured by the short-range dependent (SRD) Markovian model. The FBM process is a self-similar Gaussian process, which is therefore a model suitable for capturing the long-range dependence within traffic.

At the same time, a variety of applications of stochastic network calculus have emerged, including the evaluation of network end-to-end latency. In Zheng et al. [Zheng, Liu, Lei et al. (2013)], The performance of finite state Markov channel is analysed. The latency bound is derived based on the MGFs. Ma et al. [Ma, Chen, Li et al. (2019)] considering the characteristics of 5G architecture and used SNC to analyse the latency in URLLC. For latency-sensitive traffic, Wang et al. [Wang, Di, Jiang et al. (2017)] compared two different arrival models, the Poisson process and self-similar process, and applied the traditional scheduling strategy to MEO nodes while considering link impairment between a pair of satellites. In addition, Fidler et al. [Fidler and Rizk (2015)] used a stochastic service process to analyse the latency performance of TCP. They used using stochastic network calculus considering both backlog and latency [Lübben and Fidler (2016)].

## 3 Theoretical basis and system model

### *3.1 Network architecture*

Both the International Telecommunication Union (ITU) and 3GPP have conducted research on the convergence of satellite and terrestrial 5G networks. On the one hand, for the problem of satellite and terrestrial 5G convergence, the ITU proposed four application scenarios for satellite 5G convergence, including relay-to-station, community-backhaul, mobile-to-communication, and hybrid broadcast scenarios. On the other hand, 3GPP space-ground integrated communication related standards research is mainly carried out in two projects, TR38.811 and TR22.822. TR38.811 mainly studies the 5G new air interface standard for non-terrestrial networks, and TR22.822 mainly studies the access of 5G satellites. The architecture of Sat5G is shown in Fig. 1.

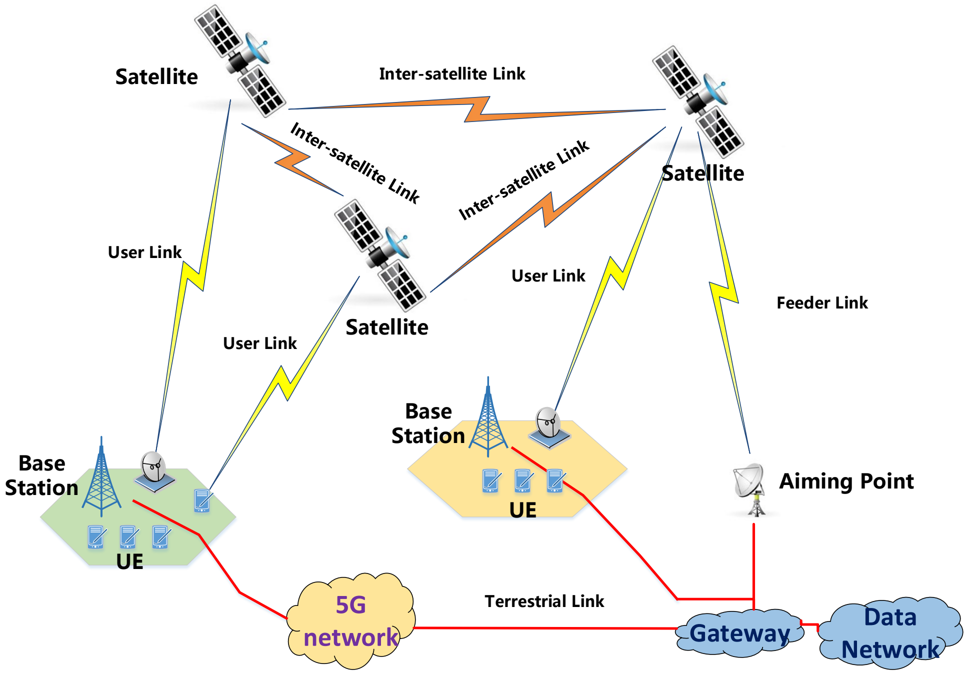

**Figure 1:** Architecture of Sat5G network

From a technical point of view, the architecture of Sat5G has both a bentpipe forwarding mode and an on-board processing mode. The two modes are different in implementation complexity and application scenarios, as illustrated in Fig. 2.

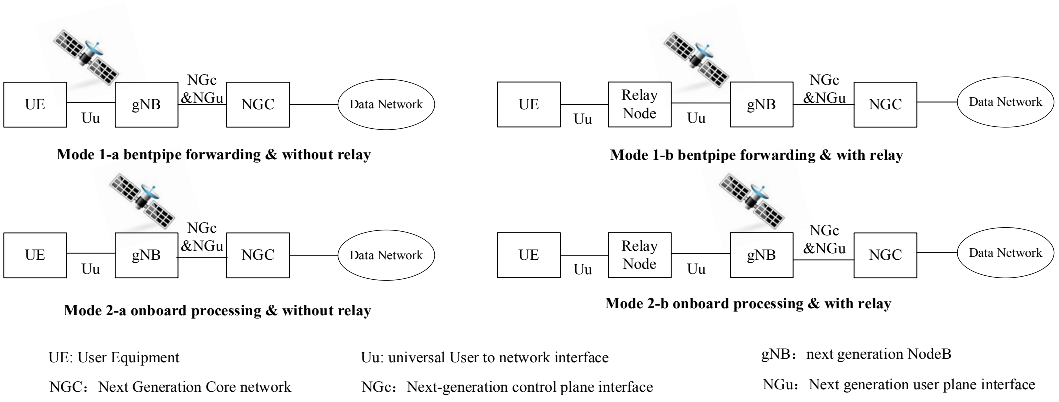
UE
Uu
gNB
NGc &NGu
NGC
Data Network
Mode 1-a bentpipe forwarding & without relay
UE
Uu
Relay Node
Uu
gNB
NGc &NGu
NGC
Data Network
Mode 1-b bentpipe forwarding & with relay
UE
Uu
gNB
NGc &NGu
NGC
Data Network
Mode 2-a onboard processing & without relay
UE
Uu
Relay Node
Uu
gNB
NGc &NGu
NGC
Data Network
Mode 2-b onboard processing & with relay
UE: User Equipment
Uu: universal User to network interface
gNB：next generation NodeB
NGC：Next Generation Core network
NGc：Next-generation control plane interface
NGu：Next generation user plane interface

**Figure 2:** Models of Sat5G network

In the long run, some or all of the ground base stations will be used. The gradual migration of functions to the satellite is a trend that can effectively reduce processing latency and improve the quality of experience (QoE) of users. Therefore, we will take mode 2-a as an example to perform end-to-end latency performance analysis.

### 3.2 Latency of Sat5G

Based on mode 2-a, the connection relationship between the various components in Sat5G is shown in Fig. 3. Here we consider the packet transmission latency from User Equipment (UE) to Data Network (DN).

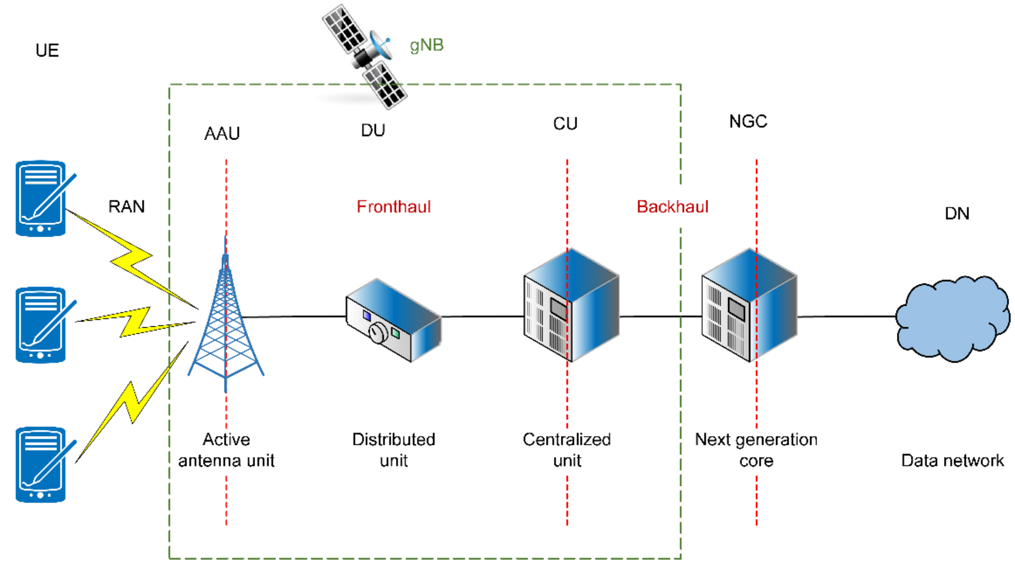

**Figure 3:** Connection relationship between the various components in Sat5G network

(1) Radio Access Network (RAN): RAN consists of User Equipment (UE) and Active Antenna Unit (AAU), AAU is part of the base-station, and UEs first accesses AAU. (2) Fronthaul: the communication process between AAU and Centralized Unit (CU) is called Fronthaul. AAU forwards the traffic to Distributed Unit (DU), when the traffic gets to DU, there are two scenarios. If both CU and DU are deployed, the traffic can reach CU immediately. Otherwise, the traffic will be sent from DU to CU. (3) Backhaul: the communication process between CU and Next Generation Core (NGC) is called Backhaul. The traffic leaves CU and goes to NGC, NGC will take some time to deal with the traffic. Finally, NGC transmits the data to Data Network (DN). The unidirectional transmission is thus completed. Consequently, the whole latency in the Sat5G system will be associated with the transmission latency in RAN, Fronthaul, Backhaul, DN, and the processing latency in NGC. We have used the calculation method in Fidler et al. [Fidler and Rizk (2015); Ma, Chen, Li et al. (2019)], which is expressed as Eq. (1).

$$T_{ETE} = T_{RAN} + T_{Fronthaul} + T_{Backhaul} + T_{NGC} + T_{DN} \tag{1}$$

In order to satisfy the QoS requirements of massive services, especially latency-sensitive services, it is essential to study $T_{ETE}$ .

### *3.3 Stochastic network calculus*

The queuing system is analysed by means of minimum plus algebra in stochastic network calculus. Let $F$ denote the set of nonnegative nondecreasing functions and $\overline{F}$ denote the set of nonnegative nonincreasing functions. A traffic flow is representing by a cumulative process. Arrival process is denoted as $A(t)$ , departure process is denoted as $D(t)$ and service process is denoted as $S(t)$, respectively. For any $0 \le s \le t$, $A(0)=0$, $A(s,t)=A(t)-A(s)$, and $A(t)$ is the cumulative arrival traffic. The same applies for $D(t)$ and $S(t)$. Referring to Fidler et al. [Fidler and Rizk (2015)], we give the following definition.

**Definition 1 (Stochastic Arrival Curve).** A traffic flow has a stochastic arrival curve $\alpha \in F$ with bounding function $f \in \overline{F}$, denoted by $A(t) \sim< f,\alpha >$, if for all $t \ge 0$ and $x \ge 0$ and $x \ge 0$ we have

$$P\left\{\sup_{0\le s\le t}\left\{A(s,t)-\alpha(t-s)\right\} > x\right\} \le f(x) \tag{2}$$

where $\alpha(\tau)$ is the arrival curve, which denotes the maximum traffic, $f(x)$ denotes the violation probability.

**Definition 2 (Stochastic Service Curve).** A system $S$ provides a stochastic service curve $\beta \in F$ with bounding function $g \in \overline{F}$, denoted by $S \sim< g,\beta >$, if for all $t \ge 0$ we have

$$P\left\{\sup_{0\le s\le t} A \otimes \beta(s) - D(s) > x\right\} \le g(x) \tag{3}$$

The symbol $\otimes$ represents the operation of cumulative min-plus convolution. where

$$A \otimes \beta(t) = \inf_{0\le s\le t}\left\{A(s)+\beta(s,t)\right\} \tag{4}$$

Here, $\beta(t)$ is the stochastic service curve which is similar to the stochastic arrival curve, The probability of generating excess traffic is constrained by the bound function $g(x)$.

Similar to Eq. (3), the departure process and is described as

$$D(t) \ge \inf_{0\le s\le t}\left\{A(s)+S(s,t)\right\}=A \otimes S(t) \tag{5}$$

From Eq. (5), The relationship among $A(t)$, $D(t)$ and $S(t)$ can be better understood.

With Eq. (2)-(5), we can now discuss the definition of the latency process and latency bound.

**Definition 3 (Latency process)** Arrival process is denoted as $A(t)$ , departure process is denoted as $D(t)$. The latency process $L(t)$ at time $t \ge 0$ can be described as

$$L(t)=\inf\left\{d \ge 0 : A(t) \le D(t+d)\right\} \tag{6}$$

$L(t)$ is the minimum value of $d$, and $d$ must satisfy the constraint that the amount of traffic arriving at time $t$ is not more than the traffic leaving at time $t+d$.

**Lemma 1 (Latency Bound).** $A(t)$ is a stochastic arrival process with a curve $\alpha \in F$ , which is bounded by function $f \in \bar{F}$ (i.e., $A \sim < f, \alpha >$ ). $S(t)$ is a stochastic service process with a curve $\beta \in F$ , which is bounded by function $g \in \bar{F}$ (i.e., $S \sim < g, \beta >$ ). Then, for all $t \geq 0$ and $x \geq 0$ , the latency bound $L(t)$ can be described as

$$P\{L(t) > h(\alpha + x, \beta)\} \leq f \otimes g(x) \tag{7}$$

where the function $h(\alpha + x, \beta)$ is the maximum value of horizontal distance between $\alpha + x$ and $\beta$ ; the expression $f \otimes g(x)$ represents the operation of cumulative min-plus convolution of functions $f$ and $g$ .

**Lemma 2 (Concatenation Property).** Consider a traffic passing through $N$ server nodes in tandem. If each node $n(n \in N)$ provides a stochastic service curve $S^n \sim < g^n, \beta^n >$ , the network guarantees a stochastic service curve $S \sim < g, \beta >$ to the traffic with

$$\begin{aligned} \beta(t) &= \beta^1 \otimes \beta^2 \otimes \cdots \otimes \beta^N(t) \\ g(x) &= g^1 \otimes g^2 \otimes \cdots \otimes g^N(x) \end{aligned} \tag{8}$$

### *3.4 Traffic model*

The traffic model is an important factor in network queuing performance analysis. A large number of WAN, LAN, MANET, Internet switches, and video and VBR traffic data were collected and analysed in detail. Finally, possible self-similarity of service flows was assumed to exist at any time and in any network environment. Here, we can give the definitions of the self-similarity of traffic and the mathematical expression of self-similarity.

**Definition 4 (Self-similarity of Traffic)** The self-similarity of network traffic means that the network flow exhibits the same burst mode at different observation time scales, that is, the burstiness of the aggregated service is maintained regardless of whether the time scale is increased or decreased.

**Definition 5 (Mathematical Expression of Self-similarity)** If an arrival process $A(s,t)$ is with stationary increments $A(s,t) \sim_{dist} A(s+\tau, t+\tau)$ , and if its deviations from a constant rate traffic $X(s,t) \sim_{dist} A(s,t) \text{-} r(t-s)$ have the self-similarity property, we call that $A(s,t)$ is self-similar with a Hurst parameter $H \in (0,1)$ .

At present, many self-similar traffic models have been proposed, such as the long correlation ON/OFF model, fractal Gaussian noise (FGN) model, fractal Brownian motion (FBM), fractal autoregressive smoothing (FARIMA) model and $M / G / \infty$ model based on the queuing process. The simplest and most commonly used model is the self-similar traffic model based on fractal Brownian motion (FBM) proposed by Norros [Akyildiz, Wang and Lin (2015)].

**Definition 6 (Fractal Brownian Motion Traffic Model)** The traffic $A(t)$ that arrives within the time period $t$ satisfies

$$A(t)=mt+\sqrt{am}Z_H(t) \tag{9}$$

where $m>0$ is the average arrival rate of the traffic, $a=\sigma^2 / m$ $a$ is the variance coefficient, $\sigma^2$ is the variance of the traffic flow within a time unit, $Z_H(t)$ is the standard fractal Brownian motion, and the self-similar parameter $H$ satisfies $H \in (0.5,1)$; then, $A(t)$ is called the fractal Brownian motion traffic model.

## 4 End to end latency calculation

### *4.1 Problem description*

As shown in Fig. 3, the data in Sat5G are transferred from UE to the data network. To satisfy the requirement of Sat5G, the end-to-end latency can be described as Eq. (10).

$$P\{T_{ETE} > D\} < \varepsilon \tag{10}$$

where $T_{ETE}$ is the whole latency in the Sat5G system and $\varepsilon$ is defined as a violation probability. Eq. (10) shows that Sat5G network transfers traffic successfully and satisfies the latency requirements.

### *4.2 Model building*

Sat5G network characteristics can be regarded as dynamic servers using the stochastic processes described in Section 3.3. Traffic from UE can be depicted as the arrival process $A(t)$. Accordingly, the service capacity of the network server node can be represented by the service process $S(t)$. Therefore, if the Sat5G is considered as a tandem system, the end-to-end latency of Sat5G should fall into the tandem characterization in stochastic network calculus.

Considering that a traffic flow from UE passes through the gNB, NGC and DN. Each network node $k$ provides a stochastic service curve $S_k \sim < g_k, \beta_k >$ to its flow. In addition, the arrival process of NGC $A_{NGC}(t)$ equals to the departure process of gNB $D_{gNB}(t)$ actually, where $A_{NGC}(t)=D_{gNB}(t)$, and so on. Based on Theorem 1 and Theorem 2, which were proven by Ma et al. [Ma, Chen, Li et al. (2019)], and Lemma 2 of this paper, we can derive the end-to-end latency bound of the Sat5G system.

**Theorem 1 (Latency Bound of Sat5G).** $A_{AAU}(t)$ is a stochastic arrival process in Sat5G with curve $\alpha_{AAU}$, i.e., $A \sim < f_{AAU}, \alpha_{AAU} >$. Here, $\alpha_{AAU} \in F$, and $f_{AAU} \in \bar{F}$. The server nodes in Sat5G provide stochastic service processes $S_{AAU}(t)$, $S_{DU}(t)$, $S_{CU}(t)$, $S_{NGC}(t)$ and $S_{DN}(t)$, i.e., $S_{AAU} \sim < g_{AAU}, \beta_{AAU} >$, $S_{DU} \sim < g_{DU}, \beta_{DU} >$, $S_{CU} \sim < g_{CU}, \beta_{CU} >$, $S_{NGC} \sim < g_{NGC}, \beta_{NGC} >$, and $S_{DN} \sim < g_{DN}, \beta_{DN} >$, respectively. The service rates $\beta_{AAU}$, $\beta_{DU}$, $\beta_{CU}$, $\beta_{NGC}$, $\beta_{DN} \in F$, $g_{AAU}$, $g_{DU}$, $g_{CU}$, $g_{NGC}$, and $g_{DN} \in \bar{F}$. Then, for all $t \geq 0$ and $x \geq 0$, the latency of the Sat5G system $T_{ETE}(t)$ is bounded by

$$P\left\{T_{ETE}\left(t\right)\geq d\right\}=P\left\{T_{ETE}\left(t\right)\geq h\left(\alpha_{AAU}+x,\beta_{all}\right)\right\}$$
$$\leq f_{AAU}\otimes g_{all}\left(x\right) \tag{11}$$

where $\beta_{all}\left(t\right)=\beta_{AAU}\otimes\beta_{DU}\otimes\beta_{CU}\otimes\beta_{NGC}\otimes\beta_{DN}\left(t\right)$ and

$g_{all}\left(x\right)=g_{AAU}\otimes g_{DU}\otimes g_{CU}\otimes g_{NGC}\otimes g_{DN}\left(x\right)$.

### 4.3 Latency calculation

Here, the latency bound of the model is calculated. With Theorem 1, we know that four key variances which are stochastic arrival processes $\alpha_{AAU}$ and $f_{AAU}$ and stochastic service processes $\beta_{all}$ and $g_{all}$ need to be derived. In particular, we can also decompose $\beta_{all}$ and $g_{all}$ to achieve more detailed results.

#### 4.3.1 Derivation of $f_{AAU}$

In Section 3.4, we provide the definition and mathematical expression of self-similarity. Fractional Brownian motion (FBM) is widely used as a self-similar flow model. In the case of $H\in\left(0.5,1\right)$, the self-similar process is long-range dependent, so the *v.b.c* stochastic arrival curve for self-similar process $\alpha\left(t\right)$ with bound function $f\left(x\right)$ is

$$P\left\{\sup_{0\leq s\leq t}\left\{A\left(s,t\right)-\alpha\left(t-s\right)\right\}>x\right\}\leq f\left(x\right) \tag{12}$$

here:

$$\begin{cases}\alpha\left(t\right)=mt+\sigma t^{H}\\ f\left(x\right)=a_{f}e^{-b_{f}x^{2(1-H)}}\end{cases}$$

where $a_{f}=e^{-\theta\theta_{1}}$ and $b_{f}=\theta$ for $\forall\theta,\theta_{1}>0$, as described by Wang et al. [Wang, Di, Jiang et al. (2017)].

Therefore, the arrival curve $\alpha_{AAU}=mt+\sigma t^{H}$, and the bound function will be $f_{AAU}\left(x\right)=a_{f}e^{-b_{f}x^{2(1-H)}}$.

#### 4.3.2 Derivation of $g_{all}$

We adopt latency-rate (LR) server to analyse scheduling algorithms in Sat5G. The feature of node service is described as $\beta\left(t\right)=R\left(t-T\right)$, where $T$ is the maximum processing delay and $R$ is the minimum service rate. For a work-conserving server with constant rate, $T=L/R$, where $L$ is the maximum value of packet size, and the bound function $g\left(x\right)=a_{g}e^{-b_{g}x}$, where $a_{g}=1$ and $b_{g}=\theta_{2}$.

Therefore, the service bound functions are $g_{AAU}\left(x\right)=a_{g_1}e^{-b_{g1}x}$ , $g_{DU}\left(x\right)=a_{g_2}e^{-b_{g2}x}$ , $g_{CU}\left(x\right)=a_{g_3}e^{-b_{g3}x}$ , $g_{NGC}\left(x\right)=a_{g_4}e^{-b_{g4}x}$ , and $g_{DN}\left(x\right)=a_{g_5}e^{-b_{g_5}x}$ . According to Theorem 1, we can obtain

$$\begin{aligned} g_{all}\left(x\right) &= g_{AAU}\otimes g_{DU}\otimes g_{CU}\otimes g_{NGC}\otimes g_{DN}\left(x\right) \\ &= \inf_{x_1+x_1+x_1+x_1+x_1=x}\sum\nolimits_{k=1}^{5}a_{g_k}e^{-b_{g_k}x_k} \end{aligned} \tag{13}$$

Applying the conclusion proposed by Sun et al. [Sun, Li and Jiang (2015)], we can hold

$$\inf_{x_1+x_2+x_3+x_4+x_5=x}\sum\nolimits_{k=1}^{5}a_{g_k}e^{-b_{g_k}x_k}=e^{\frac{-x}{\omega}}\prod_{k=1}^{5}\left(a_{g_k}b_{g_k}\omega\right)^{\frac{1}{b_{g_k}\omega}} \tag{14}$$

where $\omega=\sum\nolimits_{k=1}^{5}\frac{1}{b_{g_k}}$. With all the information we discuss above, we can obtain

$$g_{all}\left(x\right)=e^{\frac{-x}{\omega}}\prod_{k=1}^{5}\left(a_{g_k}b_{g_k}\omega\right)^{\frac{1}{b_{g_k}\omega}} \tag{15}$$

### *4.3.3 Derivation of* $\boldsymbol{f_{AAU}\otimes g_{all}}$

We have calculated $f_{AAU}\left(x\right)$ in Section 4.3.1 and $g_{all}\left(x\right)$ in Section 4.3.2, according to Eq. (11). In this section, we calculate $f_{AAU}\otimes g_{all}\left(x\right)$ to obtain the latency bound violation probability.

By combining $f_{AAU}\left(x\right)=a_f e^{-b_f x^{2(1-H)}}$ and $g_{all}\left(x\right)=e^{\frac{-x}{\omega}}\prod_{k=1}^{5}\left(a_{g_k}b_{g_k}\omega\right)^{\frac{1}{b_{g_k}\omega}}$, it can be derived that

$$\begin{aligned} & P\left\{T_{ETE}\left(t\right)\geq h\left(\alpha_{AAU}+x,\beta_{all}\right)\right\} \\ & \leq f_{AAU}\otimes g_{all}\left(x\right) \\ & = a_f e^{-b_f x^{2(1-H)}}\otimes e^{\frac{-x}{\omega}}\prod_{k=1}^{5}\left(a_{g_k}b_{g_k}\omega\right)^{\frac{1}{b_{g_k}\omega}} \end{aligned} \tag{16}$$

Further, by applying Theorem 3, which was proven by Beck [Beck (2016)], it can be derived that

$$\begin{aligned} & P\left\{T_{ETE}\left(t\right)\geq h\left(\alpha_{AAU}+x,\beta_{all}\right)\right\} \\ & \leq f_{AAU}\otimes g_{all}\left(x\right) \\ & = a_f e^{-b_f x^{2(1-H)}}\otimes e^{\frac{-x}{\omega}}\prod_{k=1}^{5}\left(a_{g_k}b_{g_k}\omega\right)^{\frac{1}{b_{g_k}\omega}} \\ & \leq a_f e^{-b_f x^{2(1-H)}}\cdot e^{\frac{-x}{\omega}}\prod_{k=1}^{5}\left(a_{g_k}b_{g_k}\omega\right)^{\frac{1}{b_{g_k}\omega}} \end{aligned} \tag{17}$$

Then, we set $D = h\left(\alpha_{AAU} + x, \beta_{all}\right) = \frac{x}{C-m}$, and make the right side of Eq. (17) equals $\varepsilon$, which means a violation probability. The relationship between $D$ and $\varepsilon$ is obtained.

$$\begin{aligned}\ln\varepsilon &= \ln\left(a_f \cdot \prod_{k=1}^{5}\left(a_{g_k} b_{g_k}\omega\right)^{\frac{1}{b_{g_k}\omega}}\right) - \left(\frac{x}{\omega} + b_f x^{2(1-H)}\right)\\ &= \ln\left(a_f \cdot \prod_{k=1}^{5}\left(a_{g_k} b_{g_k}\omega\right)^{\frac{1}{b_{g_k}\omega}}\right) - \left(\frac{C-m}{\omega}\cdot D + b_f\left(C-m\right)^{2(1-H)}\cdot D^{2(1-H)}\right)\end{aligned} \tag{18}$$

where $\omega = \sum_{k=1}^{5}\frac{1}{b_{g_k}}$, $C$ is the service rate and $m$ is the average arrival rate.

## 5 Experimental results and performance evaluation

we will evaluate the determinants of end-to-end latency of Sat5G in this section. The packet arrival process satisfies the FBM model. More simulation parameters can be found in Tab. 1.

**Table 1:** Simulation parameters

| Parameter | Value |
|---|---|
| Traffic model | FBM |
| Number of tandem servers | 5 |
| Satellite type | LEO |
| Satellite altitude | 600 km |
| Latency bound | 30 ms |

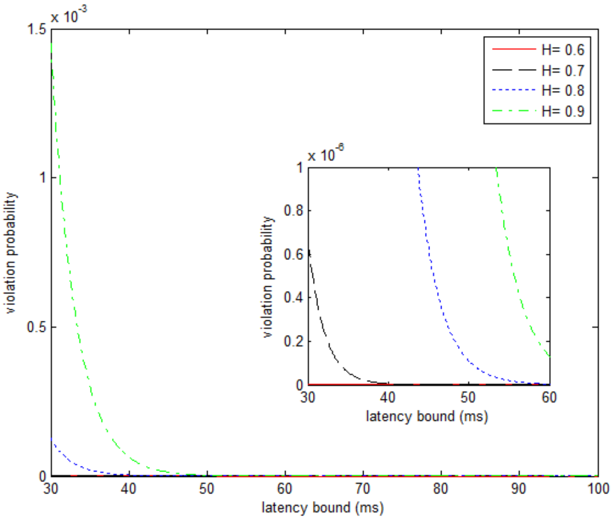

**Figure 4:** Relationship between the latency bound and violation probability under different Hurst parameters

Fig. 4 illustrates the relationship between the latency bound and violation probability under different Hurst parameters. Here, $a_f = a_{g_k} = 20$, $b_f = b_{g_k} = 1$ $(k = 1,2,3,4,5)$, $C = 3Gbps$, $m = 1.5Gbps$. As shown in the figure, on the one hand, under the same latency bound, the larger the H value, the higher the violation probability. On the other hand, under the same H value, the greater the latency bound, the lower the violation probability.

Fig. 5 illustrates the relationship between the latency bound and violation probability under different service rates. Here, $a_f = a_{g_k} = 20$, $b_f = b_{g_k} = 1$ $(k = 1,2,3,4,5)$, $m = 1.5Gbps$, and $H = 0.8$. As shown in the figure, on the one hand, under the same service rate, the greater the latency bound, the lower the violation probability is. On the other hand, under the same latency bound, the larger the service rate, the higher the violation probability is.

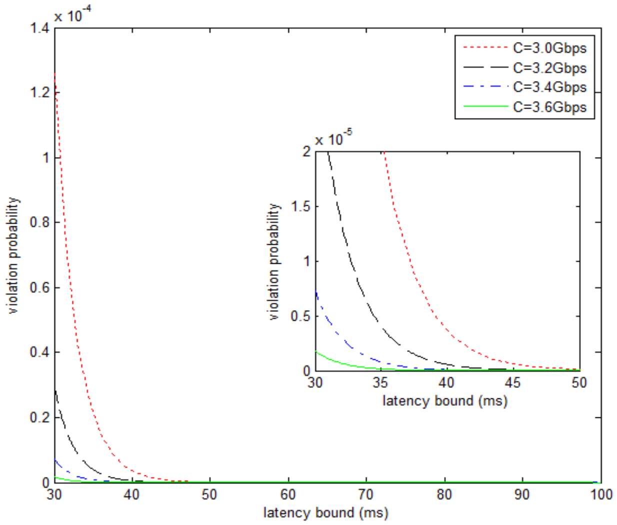

**Figure 5:** Relationship between the latency bound and violation probability under different service rates

## 6 Conclusion

In this paper, the Sat5G network is modelled as a tandem system by analysing the architecture characteristics. Through the application of SNC, the traffic model is proposed, and the performance analysis is carried out in combination with the characteristics of Sat5G network. The relationships among latency bound, service rate, Hurst parameters, violation probability and arrival rate in Sat5G network are studied. The satellite link propagation latency is considered when the simulation parameters are setting. Numerical results verify that the Hurst parameter, corresponding to the burstiness of traffic, will affect the violation probability to some extent. This also means that we need to adopt an adaptive strategy in the traffic scheduling process. The service rate is also a factor affecting latency. In the future, different scheduling strategies will be taking into account for self-similar traffic.

**Funding Statement:** This work was supported by the National Natural Science Foundation of China under Grants 61801073, 61722105, 61931004, and the Natural Science Foundation of Liaoning Province under Grant 20170540034.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

**Akyildiz, I. F.; Wang, P.; Lin, S.** (2015): SoftAir: a software defined networking architecture for 5G wireless systems. *Computer Networks*, pp. 1-18.

**Beck, M.** (2016): Towards the analysis of transient phases with stochastic network calculus. *17th International Telecommunications Network Strategy and Planning Symposium (Networks)*, pp. 164-169. IEEE, Montreal.

**Burchard, A.; Liebeherr, J.; Patek, S. D.** (2012): A calculus for end-to-end statistical service guarantees. *Computer Science*, no. 9, pp. 4105-4114.

**Chang, C. S.** (2000): *Performance Guarantees in Communication Networks*. Springer-Verlag, London.

**Cheng, Y.; Zhuang, W.; Ling, X.** (2007): FBM model based network-wide performance analysis with service differentiation. *International Conference on Heterogeneous Networking for Quality*, pp. 1-7. ACM, New York.

**Fidler, M.; Rizk, A.** (2015): A guide to the stochastic network calculus. *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 92-105.

**Giambene, G.; Kota, S.; Pillai, P.** (2018): Satellite-5G integration: a network perspective. *IEEE Network*, vol. 32, no. 5, pp. 25-31.

**Jiang, Y.; Liu, Y.** (2009): *Stochastic Network Calculus*. Springer, Germany.

**Liu, J.; Zeng, Y.; Shi, J.; Yang, Y.; Wang, R. et al.** (2019): MalDetect: a structure of encrypted malware traffic detection. *Computers, Materials and Continua*, vol. 60, no. 2, pp. 721-739.

**Liu, K. Y.; Peng, J.; Wang, J. G.; Yu, B. Y.; Liao, Z. F. et al.** (2019): A learning-based data placement framework for low latency in data center networks. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2019.2940953.

**Lokshina, I.; Zhong, H.; Lanting, C. J. M.** (2020): Self-similar teletraffic in a smart world. *Data-Centric Business and Applications*.

**Lübben, R.; Fidler, M.** (2016): Estimation method for the delay performance of closed-loop flow control with application to TCP. *IEEE INFOCOM The 35th Annual IEEE International Conference on Computer Communications*, pp. 1-9. IEEE, San Francisco.

**Ma, S.; Chen, X.; Li, Z.; Chen, Y.** (2019): Performance evaluation of URLLC in 5G based on stochastic network calculus. *Mobile Networks and Applications*, pp. 1-13.

**Mattisson, S.** (2018): An overview of 5G requirements and future wireless networks: accommodating scaling technology. *IEEE Solid-State Circuits Magazine*, vol. 10, no. 3, pp. 54-60.

**Sun, F.; Li, L.; Jiang, Y.** (2015): Impact of duty cycle on end-to-end performance in a wireless sensor network. *IEEE Wireless Communications and Networking Conference*, pp. 1906-1911. IEEE, New Orleans.

**Wang, F.; Zhang, L. L.; Zhou, S. W.; Huang, Y. Y.** (2019): Neural network-based finite-time control of quantized stochastic nonlinear systems. *Neurocomputing*, vol. 362, pp. 195-202.

**Wang, M.; Di, X.; Jiang, Y.; Li, J.; Jiang, H. et al.** (2017): End-to-End stochastic QoS performance under multi-layered satellite network. *International Conference on Space Information Networks*. Springer, Singapore.

**Yu, W.; Liang, F.; He, X.; Hatcher, W. G.; Lu, C. et al.** (2018): A survey on the edge computing for the internet of things. *IEEE Access*, vol. 6, pp. 6900-6919.

**Zheng, K.; Liu, F.; Lei, L.; Lin, C.; Jiang, Y.** (2013): Stochastic performance analysis of a wireless finite-state Markov channel. *IEEE Transactions on Wireless Communications*, vol. 12, no. 2, pp. 782-793.

**3GPP TR22.822** (2018): Study on Using Satellite Access in 5G, V0.2.0.