Research Article

Delay Bound: Fractal Traffic Passes through Network Servers

Ming Li, 1 Wei Zhao, 2 and Carlo Cattani 3

1 School of Information Science & Technology, East China Normal University, No. 500 Dongchuan Road, Shanghai 200241, China
2 Department of Computer and Information Science, University of Macau, Avenida Padre Tomas Pereira, Taipa 1356, Macau
3 Department of Mathematics, University of Salerno, Via Ponte Don Melillo, 84084 Fisciano, Italy

Correspondence should be addressed to Ming Li; ming_lihk@yahoo.com

Received 25 August 2013; Accepted 5 September 2013

Copyright © 2013 Ming Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Delay analysis plays a role in real-time systems in computer communication networks. This paper gives our results in the aspect of delay analysis of fractal traffic passing through servers. There are three contributions presented in this paper. First, we will explain the reasons why conventional theory of queuing systems ceases in the general sense when arrival traffic is fractal. Then, we will propose a concise method of delay computation for hard real-time systems as shown in this paper. Finally, the delay computation of fractal traffic passing through servers is presented.

1. Introduction

There are two categories of communications to perform the delivery of a message M from the source A to the destination B. One is in the sense of best effort. By best effort, one means that the computer communication system, which is denoted by S, does not guarantee the connection of sending M from A to B, and accordingly, the quantity of the time delay $D$ that M suffers from S may not be guaranteed, generally. User Datagram Protocol (UDP) is used for communications by best effort (Tanenbaum [1], Postel [2]). The other is in the sense of Transmission Control Protocol (TCP), which is connection oriented, implying that the connection for sending M from A to B is guaranteed ([1], Postel [3]). Guaranteed connection is the premise for guaranteeing the quantity of the time delay $D$ that M suffers from S from A to B. This is particularly the case when mission critical applications are required (Zhao and Ramamritham [4], Zhao et al. [5], Zhao and Stankovic [6], Mahapatra and Zhao [7], Rader [8], and Mahmoodi et al. [9]).

In the case of guaranteed connections, there are two types of communication systems. One is in the type of real-time systems. The other is in the type of nonreal-time ones. By real-time system, one implies that the predetermined time delay should be guaranteed (Natarajan and Zhao [10], Chakraborty and Eberspcher [11]). If the delay, which M suffers from S, exceeds the predetermined deadline of delay, one will consider that the message M is meaningless, and communication of M from A to B is taken to be a failure from a view of real-time systems.

In the field of computer communications, there are two categories of real-time systems. One is for hard-real-time systems, and the other is for soft ones. By hard-real-time systems, we mean that the time constraint, more precisely, the predetermined time delay, has to be assured. Otherwise, the communication is regarded as a failure ([4, 5, 10, 11], Buttazzo [12], Raha et al. [13], Malcolm and Zhao [14], Malcolm et al. [15], Budka et al. [16], and Liem and Mendiratta [17]). By soft-real-time systems, on the other side, we imply that the predetermined time constraint may be statistically violated with a predetermined probability ([10], Zhao and Chong [18], and Wang et al. [19]).

Recall that the time constraint mentioned above is the message delay suffering from S from A to B (Sandmann [20], Rodriguez-Pérez et al. [21], Anjum et al. [22], Papastergiou et al. [23], Panshenskov and Vakhitov [24], Kumar et al. [25], Ferrandiz et al. [26], Pin et al. [27], Flores et al. [28], Lenzini et al. [29], and Tu et al. [30]). More precisely, in the case of the Internet, this term specifically means the delay of data packets. Unless otherwise stated, this paper uses the term packet delay or delay for short.
While we mentioned above that delay serves as a key parameter in the aspect of traffic passing through servers in the field of computer networks, one may say that the delay denoted by $d$ is actually queuing time denoted by $t_q$ in terms of queuing system as illustrated in Figure 1. Queuing theory may appear complex mathematically. From the point of view of applications, however, it may be quite easy to do the performance analysis of a queuing system with the basic knowledge of statistical means and standard deviations together with a pen and a piece of paper or with a few lines of code of simple computer program (Cooper [31], Reich [32], Kendall [33], Luchak [34], Little [35], Whitt [36, 37], and Li et al. [38]). Indeed, we said so if arrival traffic $X(t)$ is Markovian as those discussed in [33–38], Jagerman [39], Doshi [40], McKenna and Mitra [41], Li and Chen [42], Brandão and Nova [43], Reiman and Simon [44], Ancker Jr. and Gafarian [45, 46], Daley [47], and Casale et al. [48].

Note that traffic of the Markovian type implies that it is light tailed. By light tail, we mean that its autocorrelation function (ACF) is exponentially decayed and so are its power spectrum density (PSD) function and probability density function (PDF) (Li and Zhao [49, 50], Li [51]). Nevertheless, traffic is heavy tailed (Loiseau et al. [52], Hernández-Campos et al. [53], Resnick [54], Takayasu et al. [55], Willinger et al. [56], Leland et al. [57], Paxson and Floyd [58], Willinger and Paxson [59], and Beran et al. [60]), which implies that the ACF of traffic is hyperbolically decayed, that is, slowly decayed (Tsybakov and Georganas [61]). To be precise, the ACF of traffic decays slowly such that it is nonintegrable, which implies long memory or long-range dependence (LRD) (Csabaï [62], Adas [63], Terdik and Gyires [64], Callado et al. [65], Owczarczuk [66], Scherrer et al. [67], Devetsikiotis and da Fonseca [68], Smith [69], Tadaki [70], Erramilli et al. [71], Karasaridis and Hatzinakos [72], Stathis and Maglaris [73], López-Ardao et al. [74], and Beran [75]). The LRD of traffic may be so strong that the variance of traffic may not exist or may be infinite ([54], Willinger et al. [76], Resnick et al. [77], López-Oliveros and Resnick [78], D'Auria and Resnick [79], and Fishman and Adan [80]). Consequently, conventional queuing theory may stop being used for analyzing queuing time or delay when arrival traffic is fractal with heavy tails or LRD such that it is of infinite variance.

Possible applications of conventional queuing theory to delay analysis are in the case of fractal traffic models with finite variance, such as fractional Brownian motion (fBm), fractional Gaussian noise (fGn); see, for example, Norros [81], Jin and Min [82], Itikhar et al. [83], Dahl and Willeman [84], Chevalier and Wein [85], Ou and Wein [86], Wein [87, 88], Harrison and Wein [89], Murata et al. [90], Bozma and Cohen [91], Haddad and Mazumdar [92], Ghosh and Weerasinghe [93], Duncan et al. [94], Li and Zhao [95], and Yue et al. [96]. However, overlarge buffer size may be required even when arrival fractal traffic is of finite variance (Albin and Samorodnitsky [97], Massoulie and Simonian [98], Heath et al. [99], Simonian and Guibert [100], Tsybakov and Georganas [101, 102], Willinger et al. [103], Kozachenko et al. [104], Carpio [105], Juneja [106], Shah and Wischik [107], and Vieira et al. [108]). The required buffer size may be so large that the value of the delay time obtained with conventional queuing theory may be impractically large for real-time systems.

The previous discussions imply that the key reason that makes the conventional queuing theory very difficult, if not impossible, to be used in the delay analysis of communication systems with fractal arrivals is the fractal properties of traffic, namely, self-similarity and LRD. Thus, fractal arrival traffic substantially challenges queuing theory of real-time systems.

As known, performance analysis of conventional queuing systems has to assume that statistical means and variances of arrival traffic exist (Cooper et al. [31–47], Pitts and Schormans [109], Stalling [110, 111], and Gibson [112]). However, generally speaking, it is inappropriate to assume that the variance of fractal traffic exists ([76–78], Li and Zhao [113], and Doukhan et al. [114]). Thus, new methodology that does not rely on statistical means and variances of arrival traffic is desired in the field of computer communication networks and real-time systems in particular.

Note that variance analysis of random functions or time series plays a key role in statistics (Bendat and Piersol [115], Gelman [116], Freedman [117], Shestik [118], Meyer [119], Lindgren and McClrath [120], and Fuller [121]) as well as conventional queuing theory [31–47], which is actually a branch of statistics (Papoulis [122], Bhat [123]). Therefore, one may see how it is significant for us to turn away from variance analysis of arrival traffic and queuing systems to treat delay analysis of fractal traffic passing through servers. Network calculus may be a promising theory to deal with delay analysis of queuing systems, irrelevant to means and variances of arrival traffic, exhibiting remarkable advances in the aspect of queuing theory.

There are two categories with respect to the theory of network calculus. One is for deterministic delay analysis of queuing systems (Le Boudec and Thiran [124], Firoiu et al. [125], Le Boudec [126], and Cruz [127]). The other is stochastic network calculus (Jiang and Liu [128], Wang et al. [129], Burchard et al. [130], Ciucu et al. [131], and Li and Knightly [132]). We should keep in mind that the theory of stochastic network calculus substantially differs from conventional queuing theory in methodology because it follows the criterion of being irrelevant to means and variances of arrival traffic.

This paper aims at presenting novel computation methods of delay of fractal traffic passing through servers without relating to the concepts of means and variances of arrival traffic.

The rest of the paper is organized as follows. We will give the brief of fractal traffic in Section 2. In Section 3, we will exhibit the result for the delay analysis of deterministic queuing theory. Section 4 presents our delay analysis of fractal traffic passing through servers. Finally, Section 5 concludes the paper.

2. Brief of Fractal Traffic

Denote by $x(t_i)$ the arrival traffic time series (traffic for short), where $t_i$ is the timestamp of the $i$th packet, where $i$ is a natural
number (Li et al. [133]). Then, $x(t_i)$ implies the data size of the $i$th packet. Since statistics of $x(t_i)$ is consistent with that of $x(t)$, we use $x(t)$ to indicate traffic for simplicity.

2.1. Non-Markovian Property and LRD. Denote by $r_{xx}(k) = E[x(i)x(i + k)]$ the ACF of $x(i)$, where $k$ is the time lag. The ACF $r_{xx}(k)$ indicates how the size of the $i$th packet correlates to that of another packet $(i + k)$ apart. If an ACF $R(k)$ is exponentially decayed, $R(k)$ may be neglected even for small $k$. For instance, suppose the ACF of a time series $B(i)$ that follows the Poisson distribution. It is given by (Bendat and Piersol [115])

$$r_{BB}(k) = \exp(-2\lambda |k|) \quad (\lambda > 0).$$

(1)

Then, in the case of $\lambda = 1$, we have

$$R_{BB}(1) \approx 0.135; \quad R_{BB}(2) \approx 0.018.$$  

(2)

Equation (2) implies that $R_{BB}(1)$ can be neglected in engineering because $x(i)$ is almost orthogonal to $x(i+1)$, letting along $R_{BB}(k)$ for $k > 1$. Therefore, $R_{BB}(k) \approx 0$ for $k > 0$. That means that $B(i)$ is memoryless. Accordingly, it is Markovian ([121, 122], Bunin [134], and Benes [135]). However, traffic $x(i)$ is non-Markovian, which is a property that distinguishes it from conventional time series, because $r_{xx}(k)$ is hyperbolically decayed in the form

$$r_{xx}(k) \sim k^{-\beta}, \quad 0 < \beta < 1, \quad k \to \infty.$$  

(3)

The above implies that

$$\sum_0^\infty r_{xx}(k) = \infty.$$  

(4)

Thus, $x(i)$ is LRD or of long memory. Consequently, it is non-Markovian (Yulmetyev et al. [136], Asgari et al. [137], van Kampen [138], Mura et al. [139], and Luczka [140]).

2.2. Property of 1/f Noise. Let $S_{xx}(\omega)$ be the PSD of $x(i)$, where $\omega$ is angular frequency. According to the Wiener theorem, which is also known as the Wiener-Khintchine theorem and sometimes as the Khinchin-Kolmogorov theorem (Robinson [141], Wiener [142, 143], Khintchine [144], and Yaglom [145]), $S_{xx}(\omega)$ is the Fourier transform of $r_{xx}(k)$. Since

$$\sum_0^\infty r_{xx}(k) = S_{xx}(\omega)|_{\omega=0} = \infty,$$  

(5)

it is easy to infer that $S_{xx}(\omega)$ is in the form

$$S_{xx}(\omega) \sim \frac{1}{\omega^\alpha}.$$  

(6)

Therefore, $x(i)$ follows 1/f noise (Mandelbrot [146, 147], Ruseckas et al. [148], Lenoir [149], Aquino et al. [150], Amir et al. [151], Carlini et al. [152, 153], Beran [154, 155], Lim and Teo [156], Eb and Lim [157], Muniandy and Lim [158], Muniandy et al. [159], Muniandy and Stanslas [160], Pinchas [161, 162], Wang and Yan [163], Bakhoun and Toma [164, 165], Yang et al. [166], Wang [167], Wornell [168], Barnes and Allan [169], Kasdin [170], and Corsini and Saletti [171]).

2.3. Self-Similarity. Traffic $x(i)$ approximately satisfies the definition of self-similarity given by

$$x(ai) = a^H x(i), \quad a > 0,$$  

(7)

where $\equiv$ denotes equality in the sense of probability distribution and $0 < H < 1$ stands for the Hurst parameter [58, 61]. In general, $H$ varies with time. Hence, traffic has the property of multifractals ([108], Vieira et al. [172], Vieira and Lee [173], Masugi and Takuma [174], Masugi [175], Veitch et al. [176], Salvador et al. [177], Nogueira et al. [178], Krishna et al. [179], Feldmann et al. [180], Ayache et al. [181], Ayache [182], Liao et al. [183], Liao [184], Carbone et al. [185, 186], Stanley and Meakin [187], Yang et al. [188], Song and Shang [189], Shang et al. [190], Kantelhardt et al. [191], Ostrowsky et al. [192], Sastry et al. [193], and Min et al. [194]).

2.4. The Hurst Parameter and Fractal Dimension. Expressing $\beta$ in (3) by $H$ yields

$$\beta = 2 - 2H.$$  

(8)

The parameter $\beta$ is the index of LRD, and $H$ is the measure of LRD ([52, 60, 75, 76, 154], Roughan et al. [195], Abry et al. [196], and Hall and Hart [197]). In the fields, people usually use $H$ instead of $\beta$ to characterize LRD of time series for dedicating the famous hydrologist Hurst [198].

We consider the local behavior of traffic $x(i)$ using its ACF $r_{xx}(k)$. For $k \to 0$, if $r_{xx}(k)$ is sufficiently smooth on $(0, \infty)$ and if

$$[r_{xx}(0) - r_{xx}(k)] \sim c|k|^\alpha, \quad 0 < \alpha \leq 2,$$  

(9)
where $c_1$ is a constant and $\alpha$ is the fractal index of $x(i)$ (Adler [199], Chan et al. [200], Davies and Hall [201], Constantine and Hall [202], Hall and Roy [203], Kent and Wood [204], Gneiting and Schlather [205], Gneiting [206], and Lim and Teo [207]), then the fractal dimension, denoted by $D_f$, of $x(i)$ is given by

$$D_f = 2 - \frac{\alpha}{2}. \quad (10)$$

Under the constraint of $0 < \alpha \leq 2$, one has

$$1 \leq D_f < 2. \quad (11)$$

2.5. Power Laws and Heavy Tails. Taqqu’s law says that the PDF of a random function $x(t)$ is in the form of a power function if it is LRD (Loiseau et al. [52], Doukhan et al. [114], Abry et al. [196], and Samorodnitsky and Taqqu [208]). Therefore, the PDF, ACF, and PSD of traffic are all in the form of power functions as can be seen from (3) and (6). When the PDF of a random function follows power laws, one says that it is heavy tailed (Adler et al. [209], Podobnik et al. [210, 211], Chen et al. [212], Xu et al. [213], Buraczewski et al. [214], Kulik and Soulier [215], Pisarenko and Rodkin [216], Resnick [217], Stanley [218], Bowers et al. [219], Eliazar and Klafter [220], Jakšić [221], Bansal et al. [222], Milojčić [223], and Pareto [224]).

Denote by $p(x)$ the PDF of $x(t)$. Then, the tail of $p(x)$ may be so heavy that its mean and variance, expressed, respectively, by (12) and (13), may not exist:

$$E[x(t)] = \int_{-\infty}^{\infty} xp(x) dx, \quad (12)$$

$$\text{Var}(x) = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx. \quad (13)$$

2.6. Remarks. Previous discussions imply the following remarks.

Remark 1. Traffic follows power laws.

Remark 2. It is LRD.

Remark 3. It is approximately self-similar.

Remark 4. It is a type of $1/f$ noise.

Remark 5. It is heavy tailed.

Remark 6. LRD is a global property of traffic, which is measured by $H$.

Remark 7. Fractal dimension $D_f$ characterizes the local self-similarity or local roughness or local smoothness of traffic.

In general, we do not talk about means and variances of traffic. Instead, we are interested in other two, namely, local self-similarity and LRD in the theory of fractal traffic.

3. Delay of Deterministic Queuing Systems

Network calculus may be applied to the delay analysis with respect to quality of service (QoS) in computer communication networks ([124–132], Cruz [225]). The issue of traffic passing through a server with respect to traffic delay can be described by Figure 2. The essential questions about it are stated as follows.

Question 1: how to model arrival traffic $A(t)$ towards assuring a predetermined delay, which is denoted by $D$, such that $d \leq D$?

Question 2: how to design a service scheme, which is denoted by $S(t)$, towards assuring a predetermined delay $D$, such that $d \leq D$?

Question 3: in order to guarantee the predetermined delay when $A(t)$ passes through $S(t)$, what is the operation among $A(t)$, $Y(t)$, and $S(t)$ such that $d \leq D$?

The answer to question 1 is about traffic modeling. The one to question 2 is about system modeling. That to the third is the relationship among the arrival $A(t)$, the server $S(t)$, and the departure traffic $Y(t) = A(t + d)$. Three answers constitute the basic of network calculus described in [124–127, 225].

3.1. Deterministic Envelope of Traffic. In order to assure a predetermined delay $D$ such that $d \leq D$, one may utilize an envelope, which is denoted by $A(t)$, of arrival traffic $x(t)$. There are two categories of envelopes of random functions. One is in the sense of statistical envelopes, and the other is in the sense of deterministic ones.

The literature regarding statistical envelopes of light-tailed random functions is rich, as they are needed in many fields of sciences and technologies, ranging from electronics engineering to ocean one; see, for example, Rice [226, 227], Velcheva et al. [228], Fang and Xie [229], Tayfun and Lo [230], Ochi and Sahinoglu [231, 232], Longuet-Higgins [233], Nigam [234], and Yang [235], just mentioning a few. Nonetheless, they cannot be taken as candidates of traffic envelopes because means and variances are essential to them [226–235].

In the society of computer science, people are interested in a type of envelopes of traffic, called bounding models of traffic (Michiel and Laevens [236]). Considering that arrival traffic has the property of $x(t) \geq 0$ (Li and Zhao [237]), following Cruz [127], and supposing that $x(t)$ is continuous for $t \geq 0$, a possible envelope in the time interval $[0, t]$ may be given by the inequality in the form

$$A(t) = \int_{0}^{t} x(t) dt \leq \sigma + \rho t. \quad (14)$$

There are two parameters in the above expression. One is $\sigma$ that characterizes the local property of $A(t)$ called the
burtiness in the field of computer networks ([124–127], [225],
McDysan [238], Kouvatsos et al. [239], and Anantharam and
Konstantopoulos [240, 241]). The other is $\rho$ that captures
the property of long-term rate of $A(t)$, citing two nice
survey papers by Mao and Panwar [242] and Fidler [243],
respectively, about (14).

As a matter of fact, on one hand, we have
\[
\lim_{t \to 0} A(t) = \lim_{t \to 0} \int_{0}^{t} x(t') dt' \leq \sigma. \tag{15}
\]
Thus, $\sigma$ characterizes the burtiness of $A(t)$. On the other
hand, one has
\[
\lim_{t \to \infty} \frac{A(t)}{t} = \lim_{t \to \infty} \frac{1}{t} \int_{0}^{t} x(t') dt' \leq \rho. \tag{16}
\]
Therefore, $\rho$ represents the long-term rate of $A(t)$. This model
of traffic is denoted by
\[
A(t) \sim (\sigma, \rho), \tag{17}
\]
with the special term “Leaky Bucket” [124–127, 225, 238, 242, 243].

The deterministic envelop of traffic, namely, $A(t)$, has the
properties remarked as follows.

Remark 8. $A(t)$ is increasing in the wide sense, implying that
$A(t_2) \geq A(t_1)$ if $t_2 \geq t_1$.

Remark 9. $A(t)$ expressed by (14) is irrelevant of statistics of
$x(t)$. Consequently, we do not need the concepts of statistical
means and variances for modeling traffic $x(t)$ by using $(\sigma, \rho)$.

Remark 10. Remark 9 is consistent in philosophy with fractal
models of traffic.

3.2. Service Curves. Denote a service curve of a server by $S(t)$; see
Figure 2. It represents a scheme of the server to allocate
enough resources, such as bandwidth, to arrival traffic $A(t)$
such that the delay $d$ does not exceed the predetermined $D$. Mathematically, $S(t)$ has the same properties of $A(t)$ as
described in Remarks 8–10. Thus, a function $S(t) \geq A(t)$ may
be a candidate of service curve (Yin and Poo [244], Pyun
et al. [245], Khanjari et al. [246], Chu et al. [247], Fulton
and Li [248], Li and Hwang [249], Lau and Li [250], Li and
Pruneski [251], Jamin et al. [252], Wu et al. [253], Chen et al.
[254], Agrawal et al. [255], Feng et al. [256], Raha et al. [257],
and Zhao and Chen [258]). Skills behind the idea of service
curves appear simple, but it is significant in the development
of linearizing nonlinear systems in general (Houssin et al.
[259], Okumura et al. [260], and Shinzawa [261]) and queuing
theory in particular [124].

3.3. Relationship among Arrival $A(t)$, Service $S(t)$, and Departure $A(t + d)$. As previously mentioned, $S(t)$ has the same
properties as those of $A(t)$. Thus, we denote by $S$ the set of
increasing functions in the wide sense. That is, $A(t), S(t) \in S$.
Let $X_1(t), X_2(t) \in S$. Then, the operation expressed by
(18) is called min-plus convolution [126, 262]
\[
X_1(t) \oplus X_2(t) = \inf_{0 \leq u \leq t} \{X_1(u) + X_2(t - u)\}. \tag{18}
\]
With the tool of min-plus convolution, referring to [124–127],
one has the relationship among $A(t), S(t)$, and $A(t + d)$ given by
\[
A(t) \ominus S(t) \leq Y(t) = A(t + d). \tag{19}
\]

3.4. Delay Computation of Single Server. The reports regarding
delay computation are rich; see, for example, [124–132,
225, 227, 233–258, 262], Raha et al. [263, 264], Ng et al. [265], Jia et
al. [266, 267], Amigo et al. [268], Lenzini et al. [269], Boggia
et al. [270], Karam and Tobagi [271], Fukš et al. [272], Wregge
et al. [273], Liebeherr et al. [274], and Golsetani [275]. In this
research, we present a novel way of delay computation, which is
stated below.

Theorem 11. Denote by $Y_{AS}(t) = A(t) \ominus S(t)$. Then, the delay
d$(t)$ that $A(t)$ suffers from $S(t)$ at time $t$ is given by
\[
d(t) \geq \frac{Y_{AS}(t) - \sigma - \rho t}{\rho}. \tag{20}
\]

Proof. According to (14) and (19), we have
\[
Y_{AS}(t) \leq A(t + d(t)) \leq \sigma + \rho (t + d(t)). \tag{21}
\]
Thus,
\[
Y_{AS}(t) \leq \sigma + \rho (t + d(t)). \tag{22}
\]
Solving $d(t)$ from the above yields (20). Thus, the theorem
results. \hfill $\square$

3.5. Guaranteed Delay of Single Server. Suppose that $D$ is
the predetermined deadline of delay. Then, the constraint of
guaranteed delay is expressed by
\[
d(t) \leq D \quad (t > 0). \tag{23}
\]
In order to achieve (23), we let
\[
\frac{Y_{AS}(t) - \sigma - \rho t}{\rho} \leq d(t) \leq D. \tag{24}
\]

Note that $Y_{AS}(t) = A(t) \ominus S(t)$. Therefore, we may design
either proper $S(t)$ or $A(t)$ or both such that (24) is satisfied.
For given $A(t)$, the following theorem gives the constraint of
$S(t)$ to assure (24).

Theorem 12. Denote the inverse of $\ominus$ by $\oplus$. Let $D$ be a
given deadline of delay. Then, (24) is satisfied if
\[
S(t) \geq A(t) \oplus \rho(D - t) - \sigma. \tag{25}
\]

Proof. Let (24) be satisfied. Then, we have
\[
\frac{Y_{AS}(t) - \sigma - \rho t}{\rho} \leq D. \tag{26}
\]
Changing the sign on the left side in the above expression produces
\[
\frac{\sigma + \rho t - Y_{AS}(t)}{\rho} \geq D. \tag{27}
\]
Therefore, one has
\[ Y_{AS}(t) = A(t) \otimes S(t) \geq \rho (D - t) - \sigma. \] (28)

From the above, using the inverse of min-plus convolution, we have
\[ S(t) \geq A(t) \oplus [\rho (D - t) - \sigma]. \] This completes the proof. \[\square\]

3.6. End-to-End Delay in Tandem Network. Consider arrival traffic \( A(t) \) that passes through \( n \) servers in series as indicated in Figure 3. In practice, there are a number of arrival traffic that concurrently join each server at its input port and there are some traffic that may leave at the output port of that server (Coulouris et al. [276]). Either the number of traffic joining a server or leaving it is uncertain. For instance, for the first server that is denoted by \( S_1(t) \), there are \( m+1 \) arrival traffic and \( j+1 \) departure ones at time \( t \). We are only interested in the arrival denoted by \( A(t) \) and the departure denoted by \( Y(t) \).

One way to find the end-to-end delay of \( A(t) \) passing through \( n \) servers in series is to find delay \( d_i(t) \) that \( A(t) \) suffers from the \( i \)-th server using Theorem 11 with the constraint stated in Theorem 12. Then, the end-to-end delay at time \( t \) is given by
\[ d(t) = \sum_{i=1}^{n} d_i(t). \] (29)

Denote \( S(t) \) the service curve of \( n \) servers in series. Then [124],
\[ S(t) = S_1(t) \otimes S_2(t) \otimes \cdots \otimes S_n(t). \] (30)

Therefore, when \( S(t) \) is designed following Theorem 12 and it is decomposed into \( n \) servers in series, (23) is guaranteed.

The discussions in the previous subsections produce the following remarks.

Remark 13. The above delay analysis and its computations do not need any information of the statistics of arrival traffic \( A(t) \).

Remark 14. The delay can be deterministically guaranteed. Hence, the deterministic queuing systems as Le Boudec and Thiran stated in [124].

The advantage described by Remarks 13 and 14 is at cost that more resources are required (Zhao [277], Davaril et al. [278]). In order to reduce the resource requirements that deterministic queuing analysis demands, stochastic network calculus is considered by computer scientists ([19, 95, 128, 130, 131, 243], Jiang et al. [279], Starobinski and Sidi [280], Ng et al. [281], Borst et al. [282], Liu et al. [283], Li et al. [284], Jiang [285], and Baccarelli et al. [286]). In what follows, we present a novel method of stochastic calculus for computing delay of fractal traffic passing through servers.

4. Novel Delay Analysis of Fractal Traffic Passing through Servers

We previously reported our bound of arrival traffic by taking into account its fractal dimension \( D_f \) and the Hurst parameter \( H \) [287]. It is in the form
\[ A(t) = \int_0^t x(u) \, du \leq r^{2D_f - 5} \sigma + a^{-H} \rho t, \] (31)
where \( r > 0, a > 0 \). Applying (31) to Theorem 11 immediately yields a novel delay computation as stated below.

Theorem 15. Denote by \( Y_{AS}(t) = A(t) \otimes S(t) \). Then, the delay \( d(t) \), which \( A(t) \) suffers from \( S(t) \) at time \( t \), is given by
\[ d(t) \geq \frac{Y_{AS}(t) - r^{2D_f - 5} \sigma + a^{-H} \rho t}{a^{-H} \rho}. \] (32)

Proof. According to (14) and (19), we obtain
\[ Y_{AS}(t) \leq A(t + d(t)) \leq r^{2D_f - 5} \sigma + a^{-H} \rho (t + d(t)). \] (33)
Solving \( d(t) \) from the above yields (32), which completes the proof. \[\square\]

Remark 16. The bandwidth regarding \( d(t) \) expressed by (34) may be generally less than that expressed by (21).

Remark 16 is true because we take into account two parameters of fractal traffic, namely, fractal dimension and the Hurst parameter. As a matter of fact,\[ (\sigma + \rho t) - (r^{2D_f - 5} \sigma + a^{-H} \rho t) = (1 - r^{2D_f - 5}) \sigma + (1 - a^{-H}) \rho \geq 0. \] (34)
The above expression implies that, for a given \( d(t) \), the bandwidth required based on Theorem 15 is less than that based on Theorem II.
Remark 17. Theorem 15 does not relate to statistical means and variances of arrival traffic.

Note that (31) represents a statistical bound of $A(t)$ because $D_f$ is a fractal parameter [199–207] and so is $H$ [198, 205–207]. $D_f$ expressed by (10) is with probability one and so is $H$ expressed by (8).

We previously mentioned several times that we are studying queuing systems irrelevant to statistical means and variances of arrival traffic because variances and or means of traffic may not exist [54, 76–80]. A common case that means and variances do not exist is for random functions that follow the Cauchy distribution (G. A. Korn and T. M. Korn [288], Rice [289], and Meyer [290]). Two papers by Field et al. [291, 292] utilized the Cauchy distribution for modeling traffic. A concise explanation of random functions without mean and variance is given by Bassingthwaighte [293]. The point, namely, irrelevant to statistical means and variances of arrival traffic, makes the queuing theory based on network calculus substantially differ from the conventional one. Considering large queue size based on conventional queuing theory when arrival is fractal, network calculus may yet be an attractive theory for guaranteeing queue size in a queuing system.

5. Conclusions

We have explained the reasons why conventional theory of queuing systems is inappropriate to be used in the delay analysis of queuing systems when arrival traffic is fractal. Then, we have given concise method of delay computation of deterministic queuing systems. Finally, we have derived the computation method of delay when arrival traffic is fractal.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under the project Grant nos. 61272402, 61072014, and 60873264 and by the 973 Plan under the project Grant nos. 2011CB302800.

References


