Analysis of Stochastic Service Guarantees in Communication Networks: A Server Model

Yuming Jiang and Peder J. Emstad

Centre for Quantifiable Quality of Service in Communication Systems *
Department of Telematics
Norwegian University of Science and Technology (NTNU), Norway
ymjiang@ieee.org, peder@q2s.ntnu.no

Abstract. Many communication networks such as wireless networks only provide stochastic service guarantees. For analyzing stochastic service guarantees, research efforts have been made in the past few years to develop stochastic network calculus, a probabilistic version of (min, +) deterministic network calculus. However, many challenges have made the development difficult. Some of them are closely related to server modeling, which include *output characterization*, *concatenation property*, *stochastic backlog guarantee*, *stochastic delay guarantee*, and *per-flow service under aggregation.* In this paper, we propose a server model, called *stochastic service curve* to facilitate stochastic service guarantee analysis. We show that with the concept of stochastic service curve, these challenges can be well addressed. In addition, we introduce *strict stochastic server* to help find the stochastic service curve of a stochastic server, which characterizes the service of the server by two stochastic processes: an ideal service process and an impairment process.

1 Introduction

Many communication networks such as wireless networks only provide stochastic service guarantees. Due to the increasing deployment and application of such networks to support real-time and multimedia applications, which require QoS guarantees, the development of an information theory for stochastic service guarantee analysis in these networks has been identified as a *grand challenge* for future networking research [22]. Towards it, *stochastic network calculus*, the probabilistic generalization of (min, +) *(deterministic) network calculus* [6][5][14], has been considered as a fundamental and important step [17].

Many challenges have made stochastic network calculus difficult. Some of them are closely related to server modeling, which include *output characterization*, *concatenation property*, *stochastic backlog guarantee*, *stochastic delay guarantee*, and *per-flow service under aggregation.* In particular, the experience from

* "Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" is appointed by The Research Council of Norway and funded by the Research Council, NTNU and UNINETT. (http://www.ntnu.no/Q2S/)

the development of the (min, +) network calculus for deterministic service guarantee analysis tells that a server model with the following properties is desired:

- (P.1) **(Output Characterization)** The output of a server can be represented using the same traffic model as the input.
- (P.2) **(Concatenation Property)** The concatenation of servers can be represented using the same server model.
- (P.3) **(Service Guarantees)** The server model can be used to derive backlog and delay guarantees.
- (P.4) **(Per-Flow Service)** The service received by a flow in an aggregate can be characterized using the same server model.

For the (deterministic) network calculus, its *service curve* server model has all these properties (P.1) - (P.4).

For stochastic service guarantee analysis, to the best of our knowledge, no server model satisfying (P.1) - (P.4) has been available in the literature. The most widely used one, which we shall call *weak stochastic service curve*, was introduced by Cruz [9]. Although authors in [18] have adopted weak stochastic service curve as the server model and derived interesting results for stochastic service guarantee analysis, the weak stochastic service curve model, while having property (P.3), generally does not support properties (P.1), (P.2) and (P.4).

The purpose of this paper is to propose a server model having properties (P.1) - (P.4). The proposed model is called *stochastic service curve.* In the paper, we first introduce the idea behind extending (deterministic) service curve to weak stochastic service curve and stochastic service curve, and discuss the relationship between them. We then prove properties (P.1) - (P.4) for stochastic service curve. In addition, to help find the stochastic service curve of a stochastic server, we introduce the concept of *strict stochastic server.* In a strict stochastic server, the service behavior of the server is characterized by two stochastic processes: an ideal service process and an impairment process. This characterization is inspired by the nature of a wireless channel: data is sent and received when the channel is in good condition, and no data is sent or received when the channel is in bad condition or impaired. We prove that a strict stochastic server under some general impairment condition has a stochastic service curve.

2 Network Model and Background

2.1 Network Model and Notation

We consider a discrete time model, where time is slotted as $0, 1, 2, \ldots$. The traffic of a flow is represented by $A(t)$ denoting the amount of traffic generated by the flow in $(0, t]$. In addition, we use $A(s, t)$ to denote the amount of traffic generated by the flow in $(s, t]$. The service provided by a server is represented similarly. Particularly, we let $S(s, t)$ be the amount of service provided by the server to its input in $(s, t]$ and use $S(t)$ to represent $S(0, t)$. By convention, we let $A(0) = 0$, $S(0) = 0$, and $A(t, t) = 0$ and $S(t, t) = 0$ for all $t \geq 0$.

When we consider the input and output of a network node, we use A to represent the input , A^* the output and S the service. Wherever necessary, we use subscripts to distinguish between different flows, and use superscripts to distinguish between different network nodes. Specifically, A_i^n and A_i^{n*} represent the input and output of flow i from node n respectively, and S_i^n the service provided to flow i by node n.

For stochastic service guarantee analysis, we shall focus on backlog and (virtual) delay, which are defined as [5] [14]:

(i) The backlog $B(t)$ at time $t(\geq 0)$ is $B(t) = A(t) - A^*(t)$;
(ii) The delay $D(t)$ at time $t(\geq 0)$ is $D(t) = \inf\{d \geq 0 : A(t) \leq A^*(t + d)\}$.

A function f is said to be wide-sense increasing if $f(s) \leq f(t)$ for all $s \leq t$ and to be wide-sense decreasing if $f(s) \geq f(t)$ for all $s \leq t$. We denote by $\mathcal{F}$ the set of wide-sense increasing functions defined for $t \geq 0$ with $f(t) = 0$ for $t < 0$; $\acute{\bar{\mathcal{F}}}$ the set of wide-sense decreasing functions defined for $t \geq 0$ with $f(t) = +\infty$ for $t < 0$. By definition, $A(t)$ and $S(t)$ belong to $\mathcal{F}$ and are additive, i.e. $A(s,u)+A(u,t) = A(s,t)$ and $S(s,u)+S(u,t) = S(s,t)$ for all $0 \leq s \leq u \leq t$.

The convolution of two functions f and g, denoted by $f \otimes g$, is defined as

$$f \otimes g(x) = \min_{0 \leq y \leq x} [f(y) + g(x - y)]. \tag{1}$$

If both f and g belong to $\mathcal{F}$, (1) is the same as the (min, +) convolution [14] and many properties of it have been proved [14]. These properties include: *closure property*, i.e. $\forall f, g \in \mathcal{F}$, $f \otimes g \in \mathcal{F}$; *commutativity*, i.e. $\forall f, g \in \mathcal{F}$, $f \otimes g = g \otimes f$; *associativity*, i.e. $\forall f, g, h \in \mathcal{F}$, $(f \otimes g) \otimes h = f \otimes (g \otimes h)$. In this paper, we shall use (1) also for functions in $\acute{\bar{\mathcal{F}}}$. Similarly, we can prove the following properties of $\otimes$ for functions in $\acute{\bar{\mathcal{F}}}$ [14]:

Lemma 1. *Basic properties of $\otimes$ in $\acute{\bar{\mathcal{F}}}$:*

- **Closure property:** $\forall f, g \in \acute{\bar{\mathcal{F}}}$, $f \otimes g \in \acute{\bar{\mathcal{F}}}$.
- **Commutativity:** $\forall f, g \in \acute{\bar{\mathcal{F}}}$, $f \otimes g = g \otimes f$.
- **Associativity:** $\forall f, g \in \acute{\bar{\mathcal{F}}}$, $(f \otimes g) \otimes h = f \otimes (g \otimes h)$.
- **Monotonicity:** $\forall f_1, f_2, g_1, g_2 \in \acute{\bar{\mathcal{F}}}$, *if* $f_1 \leq f_2$ *and* $g_1 \leq g_2$, $f_1 \otimes g_1 \leq f_2 \otimes g_2$.

We now present the definitions of *(deterministic) arrival curve* and *(deterministic) service curve* used in *(deterministic) network calculus* (e.g. [14]).

Definition 1. *A flow is said to have an arrival curve $\alpha \in \mathcal{F}$ iff for all $0 \leq s \leq t$, there holds*

$$A(s,t) \leq \alpha(t - s). \tag{2}$$

Definition 2. *A server is said to provide service curve $\beta \in \mathcal{F}$ to its input A, iff for all $t \geq 0$, its output A^* satisfies*

$$A^*(t) \geq A \otimes \beta(t). \tag{3}$$

Literature results show that service curve has all the properties (P.1) - (P.4) (e.g. see [14]). The concept of service curve, its these properties, together with the concept of arrival curve have helped the development of the (min, +) deterministic network calculus [6] [7] [8] [13] [5] [14].

2.2 Background on Stochastic Service Guarantee Analysis

Many applications, such as Internet video and audio, can tolerate some delay and loss, and may only require stochastic service guarantees. In addition, many networks such as wireless networks only provide stochastic service guarantees. Because of these, stochastic service guarantee analysis has become an increasingly important issue and attracted a lot of research attention in recent years. Towards it, *stochastic network calculus*, the probabilistic generalization of *deterministic network calculus* has been considered as an important step and several attempts have been made [23] [15] [21] [9] [3][16][18][2].

Most of these attempts assume deterministic server and have focused on the extension or generalization of arrival curve to the stochastic case. These extensions have generally resulted in two versions of stochastic arrival curve, which are called *traffic-amount-centric (t.a.c) stochastic arrival curve* and *virtual-backlog-centric (v.b.c) stochastic arrival curve* respectively [12]. A representative special case of t.a.c stochastic arrival curve is Exponentially Bounded Burstiness (EBB) [23] and its generalization Stochastically Bounded Burstiness (SBB)[21]. There are two limitations with t.a.c stochastic arrival curve, as investigated in [24] [16]. One is the difficulty in applying t.a.c stochastic arrival curve to the network case; the other is t.a.c stochastic arrival curve cannot be directly used to derive stochastic backlog and delay guarantees. To overcome these difficulties, t.a.c stochastic arrival curve needs to be converted to v.b.c stochastic arrival curve, or requires some additional restriction on traffic (e.g. [16]). In contrast, v.b.c stochastic arrival curve does not have these limitations. A representative special case of v.b.c stochastic arrival curve is generalized Stochastically Bounded Burstiness (gSBB)[24] (also called stochastic smoothness constraint in [9]). Under deterministic server assumption, v.b.c stochastic arrival curve has been used to analyze stochastic backlog and delay guarantees in both single node and network cases [18][12]. In addition, it is shown in [12] that many well-known types of traffic can be readily represented using v.b.c stochastic arrival curve. In this paper, we adopt v.b.c stochastic arrival curve as the traffic model.

Definition 3. *A flow is said to have a* virtual-backlog-centric (v.b.c) stochastic arrival curve $\alpha \in \mathcal{F}$ *with bounding function* $f \in \bar{\mathcal{F}}$, *denoted by* $A \sim_{vb} \langle f, \alpha \rangle$, *iff for all* $t \geq 0$ *and all* $x \geq 0$, *there holds*

$$P\{\max_{0 \leq s \leq t} \{A(s,t) - \alpha(t-s)\} > x\} \leq f(x). \tag{4}$$

The following result introduced in [20] [5] [12] can be used to find the v.b.c stochastic arrival curve of a flow:

Lemma 2. *Suppose* $a(t) \equiv A(t) - A(t-1)$ *is stationary and ergodic. Then, if* $E\{a(1)\} < r$, *there holds, for all* $t \geq 0$ *and* $x \geq 0$,

$$P\{W(t;r) > x\} \leq P\{W(t+1;r) > x\} \leq \cdots \leq P\{W(\infty;r) > x\}, \tag{5}$$

where $W(t;r) \equiv \max_{0 \leq s \leq t}[A(s,t) - r(t-s)]$ *and* $W(\infty;r)$ *denotes the steady state of* $W(t;r)$ *as* $t \to \infty$.

Note that $\max_{0 \le s \le t}[A(s,t) - r(t-s)]$ can be interpreted as the queue length at time t of a virtual single server queue (SSQ) with service rate r fed with the same traffic [24][12]. Then, the monotonicity property implies that if the traffic of the flow is stationary and ergodic, the steady-state queue length distribution of the SSQ can be used as the bounding function $f(x)$. Consequently, if the steady-state queue length distribution of a flow in a SSQ is known, then it has a v.b.c stochastic arrival curve $A \sim_{vb} \langle f, \alpha \rangle$ with $f(x) = P\{q > x\}$, the steady state compliment queue length distribution. With these, many well-known types of traffic, including Poisson, Markov Modulated Process, effective bandwidth, $\alpha-$stable, etc., can be shown to have v.b.c stochastic arrival curves [12].

While many attempts have been made for stochastic traffic modeling and analysis as discussed above, only a few have considered stochastic server and stochastic service guarantee in networks of such servers [15][9][18][16]. Essentially, the stochastic server models proposed or used in these attempts can be mapped to the following model, which we call *weak stochastic service curve* and is based on a stochastic server model used in [9]:

Definition 4. *A server S is said to provide a* weak stochastic service curve *$\beta \in \mathcal{F}$ with bounding function $g \in \bar{\mathcal{F}}$, denoted by $S \sim_{ws} \langle g, \beta \rangle$, iff for all $t \ge 0$ and all $x \ge 0$, there holds*

$$P\{A \otimes \beta(t) - A^*(t) > x\} \le g(x). \tag{6}$$

Comparing Definition 4 with Definition 2, it is clear that weak stochastic service curve is an intuitively simple generalization of (deterministic) service curve. One can easily verify that if a server has a deterministic service curve β, it has a weak stochastic service curve $S \sim_{ws} \langle 0, \beta \rangle$. In addition, the Exponentially Bounded Fluctuation (EBF) model proposed in [15] is a special case of weak stochastic service curve with an exponential form bounding function. The stochastic server model *effective service curve* used in [16] can also be verified to be a special case of weak stochastic service curve.

In [9], [18] and [16], some results have been derived based on weak stochastic service curve. The difference between them is that while [16] uses t.a.c stochastic arrive curve as the traffic model, [18] and [9] use v.b.c stochastic arrive curve. In addition to backlog and delay at a single node, [18] has considered the network case. Nevertheless, weak stochastic service curve generally does not have properties (P.1), (P.2) and (P.4) as to be explained in the remarks in the next section.

3 Stochastic Service Curve

In this section, we first investigate the duality principle of service curve, which is the idea behind the generalization of service curve to its probabilistic versions. We then introduce a new stochastic server model, called *stochastic service curve*. Stochastic service guarantee analysis is further conducted based on the new server model. Particularly, properties (P.1) - (P.4) are proved for stochastic service curve.

3.1 Definition of Stochastic Service Curve

The following result presents the duality principle of service curve. Its proof is trivial and can be found from [11].

Lemma 3. *For any constant* $\sigma \geq 0$*,* $A \otimes \beta(t) - A^*(t) \leq \sigma$ *for all* $t \geq 0$*, if and only if* $\max_{0 \leq s \leq t}\{A \otimes \beta(s) - A^*(s)\} \leq \sigma$ *for all* $t \geq 0$*, where* $\beta \in \mathcal{F}$*.*

By letting $\sigma = 0$, the first part of Lemma 3 defines a service curve β. In this case, Lemma 3 implies that if a server has service curve β or $A^* \geq A \otimes \beta(t)$, then there holds $\max_{0 \leq s \leq t}[A \otimes \beta(s) - A^*(s)] \leq 0$ and vice versa. It is in this sense we call Lemma 3 the *duality principle of service curve.*

Comparing the first part of Lemma 3 with Definition 4, one can find that the former is the basis for generalizing service curve to weak stochastic service curve. Based on the second part of the duality principle of service curve, we define the following stochastic server model, called *stochastic service curve* [1]:

Definition 5. *A server* S *is said to provide a* stochastic service curve $\beta \in \mathcal{F}$ *with bounding function* $g \in \bar{\mathcal{F}}$*, denoted by* $S \sim_{sc} \langle g, \beta \rangle$*, iff for all* $t \geq 0$ *and all* $x \geq 0$*, there holds*

$$P\{\max_{0 \leq s \leq t}[A \otimes \beta(s) - A^*(s)] > x\} \leq g(x). \tag{7}$$

Stochastic service curve implies weak stochastic service curve, since we always have $A \otimes \beta(t) - A^*(t) \leq \max_{0 \leq s \leq t}[A \otimes \beta(s) - A^*(s)]$ for all $t \geq 0$. Formally,

Lemma 4. *If a server provides stochastic service curve* $S \sim_{sc} \langle g, \beta \rangle$*, then it also provides weak stochastic service curve* $S \sim_{ws} \langle g, \beta \rangle$*.*

3.2 Properties of Stochastic Service Curve

We now study Properties (P.1) - (P.4) for stochastic service curve. For proving these properties, we need the following result. For random variables X and Y, there holds

$$P\{X + Y > x\} \leq f_X \otimes f_Y(x) \tag{8}$$

where $f_X(x) = P\{X > x\}$ and $f_Y(x) = P\{Y > x\}$. The proof of (8) can be found from the literature (e.g. [9][19][2]). With the monotonicity property of $\otimes$, if $P\{X > x\} \leq f(x)$ and $P\{Y > x\} \leq g(x)$, we get from (8) that

$$P\{X + Y > x\} \leq f \otimes g(x). \tag{9}$$

[1] In [1], *service curve with loss* is defined. It should be noticed that this definition is different from Definitions 4, 5 and 6 here. In a service curve with loss network element, packets are dropped if their deadlines assigned via the (deterministic) service curve are not met. However, in a network element with weak stochastic service curve or stochastic service curve or strict stochastic service curve, packets are allowed to violate their deadlines if they would be given such deadlines via the corresponding (deterministic) service curve.

Theorem 1. (Output) *Consider a server fed with a flow. If the server provides stochastic service curve $S \sim_{sc} \langle g, \beta \rangle$ to the flow and the flow has v.b.c stochastic arrival curve $A \sim_{vb} \langle f, \alpha \rangle$, then the output of the flow from the server has a v.b.c stochastic arrival curve $A^* \sim_{vb} \langle f^*, \alpha^* \rangle$ with $\alpha^*(t) = \max_{s \geq 0}[\alpha(t+s) - \beta(s)]$ and $f^*(x) = f \otimes g(x)$.*

Proof. Note that the output up to time t cannot exceed the input in $[0,t]$, or $A^*(t) \leq A(t)$. We now have,

$$\begin{aligned}
& \max_{0 \leq s \leq t} [A^*(s,t) - \alpha^*(t-s)] \\
= & \max_{0 \leq s \leq t} [A^*(t) - A^*(s) - \alpha^*(t-s)] \leq \max_{0 \leq s \leq t} [A(t) - A^*(s) - \alpha^*(t-s)] \\
= & \max_{0 \leq s \leq t} [A(t) - A \otimes \beta(s) - \alpha^*(t-s) + A \otimes \beta(s) - A^*(s)] \\
\leq & \max_{0 \leq s \leq t} [A(t) - A \otimes \beta(s) - \alpha^*(t-s)] + \max_{0 \leq s \leq t} [A \otimes \beta(s) - A^*(s)] \qquad (10)
\end{aligned}$$

in which,

$$\begin{aligned}
& \max_{0 \leq s \leq t} [A(t) - A \otimes \beta(s) - \alpha^*(t-s)] \\
= & \max_{0 \leq s \leq t} [A(t) - \min_{0 \leq u \leq s} [A(u) + \beta(s-u)] - \alpha^*(t-s)] \\
= & \max_{0 \leq s \leq t} \max_{0 \leq u \leq s} [A(t) - A(u) - \beta(s-u) - \alpha^*(t-s)] \leq \max_{0 \leq s \leq t} \max_{0 \leq u \leq s} [A(u,t) - \alpha(t-u)] \\
& \qquad (11) \\
= & \max_{0 \leq u \leq t} \max_{u \leq s \leq t} [A(u,t) - \alpha(t-u)] = \max_{0 \leq u \leq t} [A(u,t) - \alpha(t-u)] \qquad (12)
\end{aligned}$$

where the step (11) follows because $\alpha^*(t-s) = \max_{\tau \geq 0}[\alpha(t-s+\tau) - \beta(\tau)] \geq \alpha(t-u) - \beta(s-u)$.

Applying (12) to (10), since $S \sim_{sc} \langle g, \beta \rangle$ and $A \sim_{vb} \langle f, \alpha \rangle$, or $P\{\max_{0 \leq s \leq t}[A \otimes \beta(s) - A^*(s)] > x\} \leq g(x)$ and $P\{\max_{0 \leq u \leq t}[A(u,t) - \alpha(t-u)] > x\} \leq f(x)$, we then get from (9), $P\{\max_{0 \leq s \leq t}[A^*(s,t) - \alpha(t-s)] + \min_{s \geq 0}[\beta(s) - \alpha(s)] > x\} \leq f \otimes g(x)$, from which the theorem follows.

Remarks: (i) Note that in (10), its right hand side has a term $\max_{0 \leq s \leq t}[A \otimes \beta(s) - A^*(s)]$. If the server only has weak stochastic service curve, what is known is $P\{A \otimes \beta(s) - A^*(s) > x\} \leq g(x)$ and it is hard to find $P\{\max_{0 \leq s \leq t}[A \otimes \beta(s) - A^*(s)] > x\}$ that is critical for proving (P.1). This explains why weak stochastic service curve does not have property (P.1), if the input is modeled with v.b.c stochastic arrival curve.

(ii) If α is subadditive, following similar steps as in the above proof, we can prove that the output also has v.b.c stochastic arrival curve $A^* \sim_{vb} \langle f \otimes g(x + \min_{t \geq 0}[\beta(t) - \alpha(t)], \alpha \rangle$.

Theorem 2. (Concatenation) *Consider a flow passing through a network of N nodes in tandem. If each node $n (= 1, 2, \ldots, N)$ provides stochastic service curve $S^n \sim_{sc} \langle g^n, \beta^n \rangle$ to its input, then the network guarantees to the flow a stochastic service curve $S^* \sim_{sc} \langle g^*, \beta^* \rangle$ with $\beta^*(t) = \beta^1 \otimes \beta^2 \otimes \cdots \otimes \beta^N(t)$ and $g^*(x) = g^1 \otimes g^2 \otimes \cdots \otimes g^N(x)$.*

Proof. We shall only prove the two-node case, from which the proof can be easily extended to the N-node case. For the two-node case, the output of the first node is the input of the second node, so, $A^{1*}(t) = A^2(t)$. In addition, the input of the network is the input to the first node, or $A(t) = A^1(t)$, and the output of the network is the same as the output of the second node, or $A^* = A^{2*}$, where $A(t)$ and A^* denotes the input to and output from the network respectively. We then have,

$$\max_{0\le s\le t}[A \otimes \beta^1 \otimes \beta^2(s) - A^*(s)] = \max_{0\le s\le t}[(A^1 \otimes \beta^1) \otimes \beta^2(s) - A^{2*}(s)]. \tag{13}$$

Now let us consider any s, $(0 \le s \le t)$, for which we get,

$$\begin{aligned}
&[(A^1 \otimes \beta^1) \otimes \beta^2(s) - A^{2*}(s)] - X^1(t) - X^2(t)\\
&\le (A^1 \otimes \beta^1) \otimes \beta^2(s) - A^{2*}(s) - X^1(s) - X^2(s)\\
&\le \min_{0\le u\le s}[A^1 \otimes \beta^1(u) + \beta^2(s-u)] - \max_{0\le u\le s}[A^1 \otimes \beta^1(u) - A^2(u)] - \max_{0\le u\le s}[A^2 \otimes \beta^2(u)]\\
&\le \min_{0\le u\le s}[(A^1 \otimes \beta^1(u) + \beta^2(s-u)) - (A^1 \otimes \beta^1(u) - A^2(u))] - \max_{0\le u\le s}[A^2 \otimes \beta^2(u)]\\
&= \min_{0\le u\le s}[A^2(u) + \beta^2(s-u)] - \max_{0\le u\le s}[A^2 \otimes \beta^2(u)]\\
&= A^2 \otimes \beta^2(s) - \max_{0\le u\le s}[A^2 \otimes \beta^2(u)] \le 0.
\end{aligned} \tag{14}$$

where $X^i(t) \equiv \max_{0\le u\le t}[A^i \otimes \beta^i(u) - A^{i*}(u)]$, $i = 1, 2$.

Applying (14) to (13), we obtain

$$\max_{0\le s\le t}[A\otimes\beta^1\otimes\beta^2(s)-A^*(s)] \le \max_{0\le u\le t}[A^1\otimes\beta^1(u)-A^{1*}(u)]+\max_{0\le u\le t}[A^2\otimes\beta^2(u)-A^{2*}(u)], \tag{15}$$

with which, since both nodes provide stochastic service curve to their input, the theorem follows from (9) and the definition of stochastic service curve.

Remark: In deriving (14), we have proved $[(A^1 \otimes \beta^1) \otimes \beta^2(s) - A^{2*}(s)] \le \max_{0\le u\le s}[A^1 \otimes \beta^1(u) - A^{1*}(u)] + \max_{0\le u\le s}[A^2 \otimes \beta^2(u) - A^{2*}(u)]$ for all $s \ge 0$. However, if we want to prove concatenation property for weak stochastic service curve, we need to prove $[(A^1 \otimes \beta^1) \otimes \beta^2(s) - A^{2*}(s)] \le [A^1 \otimes \beta^1(s) - A^{1*}(s)] + [A^2 \otimes \beta^2(s) - A^{2*}(s)]$ for all $s \ge 0$, which is difficult to obtain and does not hold in general. This explains why weak stochastic service curve does not have property (P.2).

The following lemma presents stochastic backlog and stochastic delay guarantees, or property (P.3), provided by a server with weak stochastic service curve. Its proof can be found from [11] and similar results can be found from the literature (e.g. see [9] [18]). Since stochastic service curve implies weak stochastic service curve as stated by Lemma 4, Theorem 3 follows from Lemma 5.

Lemma 5. *Consider a server fed with a flow. If the server provides weak stochastic service curve $S \sim_{ws} \langle g, \beta\rangle$ to the flow and the flow has v.b.c stochastic arrival curve $A \sim_{vb} \langle f, \alpha\rangle$, then, for all $t \ge 0$ and all $x \ge 0$, (1) $P\{B(t) > x\} \le f \otimes g(x + \min_{t\ge0}[\beta(t) - \alpha(t)])$, and (2) $P\{D(t) > x\} \le f \otimes g(\min_{s\ge -x}[\beta(s + x) - \alpha(s)])$.*

Theorem 3. (Service Guarantees) *Consider a server fed with a flow. If the server provides stochastic service curve $S \sim_{sc} \langle g, \beta \rangle$ to the flow and the flow has v.b.c stochastic arrival curve $A \sim_{vb} \langle f, \alpha \rangle$, then*

- *The backlog $B(t)$ of the flow in the server at time t satisfies: for all $t \geq 0$ and all $x \geq 0$, $P\{B(t) > x\} \leq f \otimes g(x + \min_{s \geq 0}[\beta(s) - \alpha(s)])$;*
- *The delay $D(t)$ of the flow in the server at time t satisfies: for all $t \geq 0$ and all $x \geq 0$, $P\{D(t) > x\} \leq f \otimes g(\min_{s \geq -x}[\beta(s + x) - \alpha(s)])$.*

Finally, the following theorem presents per-flow service under aggregation or property (P.4) for stochastic service curve.

Theorem 4. (Per-Flow Service) *Consider a server fed with a flow A that is the aggregation of two constituent flows A_1 and A_2. Suppose the server provides stochastic service curve $S \sim_{sc} \langle g, \beta \rangle$ to the aggregate flow A.*

- *If flow A_2 has (deterministic) arrival curve α_2, then the server guarantees stochastic service curve $S_1 \sim_{sc} \langle g_1, \beta_1 \rangle$ to flow A_1, where, $g_1(x) = g(x); \quad \beta_1(t) = \beta(t) - \alpha_2(t)$.*
- *If flow A_2 has v.b.c stochastic arrival curve $A_2 \sim_{vb} \langle f_2, \alpha_2 \rangle$, then the server guarantees to flow A_1 weak stochastic service curve $S_1 \sim_{ws} \langle g'_1, \beta'_1 \rangle$, where, $g'_1(x) = g \otimes f_2(x); \quad \beta'_1(t) = \beta(t) - \alpha_2(t)$.*

Proof. For the output, there holds $A^*(t) = A_1^*(t) + A_2^*(t)$. In addition, we have $A^*(t) \leq A(t)$, $A_1^*(t) \leq A_1(t)$, and $A_2^*(t) \leq A_2(t)$. We now have for any $s \geq 0$,

$$\begin{aligned} A_1 \otimes (\beta - \alpha_2)(s) - A_1^*(s) &= \min_{0 \leq u \leq s} [A(u) + (\beta - \alpha_2)(s - u) - A_2(u)] - A^*(s) + A_2^*(s) \\ &\leq [A \otimes \beta(s) - A^*(s)] + A_2(s) - \min_{0 \leq u \leq s} [A_2(u) + \alpha_2(s - u)] \\ &= [A \otimes \beta(s) - A^*(s)] + \max_{0 \leq u \leq s} [A_2(u, s) - \alpha_2(s - u)]. \end{aligned} \tag{16}$$

For the first part, with (16), we have

$$\max_{0 \leq s \leq t} [A_1 \otimes (\beta - \alpha_2)(s) - A_1^*(s)] \leq \max_{0 \leq s \leq t} [A \otimes \beta(s) - A^*(s)] + \max_{0 \leq s \leq t} \max_{0 \leq u \leq s} [A_2(u, s) - \alpha_2(s - u)]. \tag{17}$$

Since A_2 has deterministic arrival curve α_2 and $A_2(u, s) \leq \alpha_2(s - u)$ for all $0 \leq u \leq s$, we hence have $\max_{0 \leq s \leq t} \max_{0 \leq u \leq s} [A_2(u, s) - \alpha_2(s - u)] \leq 0$, with which, $\max_{0 \leq s \leq t} [A_1 \otimes (\beta - \alpha_2)(s) - A_1^*(s)] \leq \max_{0 \leq s \leq t} [A \otimes \beta(s) - A^*(s)]$. Then, the first part follows from the definition of stochastic service curve.

For the second part, we further get from (16) that

$$A_1 \otimes (\beta - \alpha_2)(s) - A_1^*(s) \leq \max_{0 \leq u \leq s} [A \otimes \beta(s) - A^*(s)] + \max_{0 \leq u \leq s} [A_2(u, s) - \alpha_2(s - u)] \tag{18}$$

with which, $S \sim_{sc} \langle g, \beta \rangle$ and $A_2 \sim_{vb} \langle f_2, \alpha_2 \rangle$, the second part follows from (9).

Remark: Theorem 4 proves that a flow in an aggregate receives a stochastic service curve from a stochastic server when other flows in the aggregate have deterministic arrival curve. If other flows in the aggregate only have v.b.c stochastic arrival curve, what has been proved is that the flow only receives weak stochastic service curve. The difficulty in proving stochastic service curve for the flow can be found from (17), where $P\{\max_{0\leq s\leq t}\max_{0\leq u\leq s}[A_2(u,s)-\alpha_2(s-u)]>x\}$ is difficult to obtain from the given assumptions for the second part of Theorem 4. Nevertheless, we believe Theorem 4 can be readily used for stochastic service guarantee analysis in many network scenarios. One is the single node case. With Theorem 4 and Lemma 5, per-flow stochastic backlog and delay guarantees can be derived for the single node case. Another scenario is the analysis of Differentiated Services (DiffServ) in wireless networks. Under DiffServ, the Expedited Forwarding (EF) class is deterministically regulated and usually put at the highest priority level. In this scenario, Theorem 4 sheds some light on deriving stochastic service curve and stochastic service guarantees for DiffServ Assured Forwarding (AF) that is given lower priority than EF.

4 Strict Stochastic Server

In this section, we introduce *strict stochastic server* to help find the stochastic service curve of a stochastic server, which is inspired by an intuition.

In wireless networks, the behavior of a wireless channel is most simply revealed by the following intuition. The channel operates in two states: "good" and "bad". If the channel condition is good, data can be sent from the sender to the receiver at the full rate of the channel; if the condition is bad, no data can be sent. The bad channel condition has various causes such as noise, fading, contention, etc, which in all we shall call *impairment*.

Inspired by the above intuition, we use two stochastic processes to characterize the behavior of a stochastic server. These two processes are (1) an *ideal service process* $\hat{S}$ and (2) an *impairment process* I. Here, $\hat{S}(s,t)$ denotes the amount of service that the server would have delivered in interval $(s,t]$ if there had been no service impairment, and $I(s,t)$ denotes the amount of service, called *impaired service*, that cannot be delivered in the interval to the input due to some impairment to the server. Particularly, we have that the actually delivered service to the input satisfies, for all $t\geq 0$,

$$S(t)=\hat{S}(t)-I(t), \tag{19}$$

where $\hat{S}(t)\equiv\hat{S}(0,t)$ and $I(t)\equiv I(0,t)$ with $\hat{S}(0)=0$ and $I(0)=0$ by convention. It is clear that $\hat{S}, I$ are in $\mathcal{F}$ and additive.

We now define *strict stochastic server* as follows:

Definition 6. *A server S is said to be a* strict stochastic server *providing* strict stochastic service curve $\hat{\beta}(\cdot)\in\mathcal{F}$ *with impairment process I to a flow iff during any backlogged period $(s,t]$, the output $A^*(s,t)$ of the flow from the server satisfies*

$$A^*(s,t)\geq\hat{\beta}(t-s)-I(s,t).$$

In the rest, we assume $\hat{\beta}$ is additive and has the form of $\hat{\beta}(t)=\hat{r}t$. In addition, we assume the impairment process $I(t)$ is $(\sigma(\theta),\rho(\theta))$-upper constrained, a model that was initially used in [4] to characterize stochastic behavior of traffic, whose definition is as follows:

Definition 7. *A stochastic sequence I, $I \equiv \{I(t), t = 0, 1, 2, \ldots\}$ with $I(0) = 0$, is said to be $(\sigma(\theta), \rho(\theta))$-upper constrained (for some $\theta > 0$), iff for all $0 \le s \le t$*

$$\frac{1}{\theta} log E e^{\theta(I(t)-I(s))} \le \rho(\theta)(t-s) + \sigma(\theta). \tag{20}$$

The following result shows that if the impairment process $I(t)$ is $(\sigma(\theta), \rho(\theta))$-upper constrained, a strict stochastic server has a stochastic service curve. Due to space limitation, the proof is omitted and can be found from [11].

Theorem 5. *Consider a strict stochastic server providing strict stochastic service curve $\hat{\beta}(t) = \hat{r} \cdot t$ with impairment process I to a flow. Suppose I is $(\sigma(\theta), \rho(\theta))$-upper constrained. Then, the server provides to the flow a stochastic service curve $S \sim_{sc} \langle \beta, g \rangle$, where $\beta(t) = p\hat{r} \cdot t$ and $g(x) = \frac{e^{\theta\sigma(\theta)}}{(1-e^{\theta(\rho(\theta)-(1-p)\hat{r})})^2} e^{-\theta x}$ with any p, $(0 \le p < 1)$, satisfying $(1-p)\hat{r} > \rho(\theta)$.*

Remark: The definition of strict stochastic server is based on the intuition on the behavior of a wireless channel, which provides a simple approach to characterize this behavior. Theorem 5 proves the stochastic service curve of a strict stochastic server, with which and the analysis in the previous section, stochastic service guarantees can be derived for networks of strict stochastic servers. In addition, in [4] [5], many well known processes such as Markov Modulated Processes have been proved to be $(\sigma(\theta), \rho(\theta))$-upper constrained. Note that these processes have also been used in the literature for characterizing a wireless channel (e.g. [9] [10]). We hence believe our results are useful for stochastic service guarantee analysis in such networks.

5 Conclusion

In this paper, we introduced a new server model, called *stochastic service curve*, for stochastic service guarantee analysis. We have proved that stochastic service curve has properties (P.1)-(P.4), which are crucial for stochastic service guarantee analysis and the development of stochastic network calculus. In addition, we have proposed the concept of *strict stochastic server* to help find the stochastic service curve of a stochastic server. In a strict stochastic server, the service is characterized by two stochastic processes: an ideal service process and an impairment process. The impairment process provides a simple approach to model the impairment experienced by a server, which is typical in wireless networks.

While property (P.4), i.e. per-flow service under aggregation, has been proved for stochastic service curve, it is based on the assumption that other flows in the aggregate are deterministically upper-bounded. It would be interesting to prove stochastic service curve for property (P.4), when the other flows in the aggregate are only stochastically upper-bounded. Future work could hence be conducted to design traffic and server models to have properties (P.1)-(P.4) without additional assumptions on traffic or server.

References

1. S. Ayyorgun and R. L. Cruz. A composable service model with loss and a scheduling algorithm. In *Proc. IEEE INFOCOM'04*, 2004.
2. S. Ayyorgun and W. Feng. A systematic approach for providing end-to-end probabilistic QoS guarantees. In *Proc. IEEE IC3N'04*, 2004.
3. A. Burchard, J. Liebeherr, and S. D. Patek. A calculus for end-to-end statistical service guarantees. Technical report, CS-2001-19, University of Virginia, 2002.
4. C.-S. Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Trans. Automatic Control*, 39(5):913–931, May 1994.
5. C.-S. Chang. *Performance Guarantees in Communication Networks.* Springer-Verlag, 2000.
6. R. L. Cruz. A calculus for network delay, part I: network elements in isolation. *IEEE Trans. Information Theory*, 37(1):114–131, Jan. 1991.
7. R. L. Cruz. A calculus for network delay, part II: network analysis. *IEEE Trans. Information Theory*, 37(1):132–141, Jan. 1991.
8. R. L. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE JSAC*, 13(6):1048–1056, Aug. 1995.
9. R. L. Cruz. Quality of service management in integrated services networks. In *Proc. 1st Semi-Annual Research Review, CWC, UCSD*, June 1996.
10. M. Hassan, M. M. Krunz, and I. Matta. Markov-based channel characterization for tractable performance analysis in wireless packet networks. *IEEE Trans. Wireless Communications*, 3(3):821–831, May 2004.
11. Y. Jiang and P. J. Emstad. Analysis of stochastic service guarantees in communication networks: A server model. Technical report, Q2S, NTNU, April 2005.
12. Y. Jiang and P. J. Emstad. Analysis of stochastic service guarantees in communication networks: A traffic model. Technical report, Q2S, NTNU, Feb. 2005.
13. J.-Y. Le Boudec. Application of network calculus to guaranteed service networks. *IEEE Trans. Information Theory*, 44(3):1087–1096, May 1998.
14. J.-Y. Le Boudec and P. Thiran. *Network Calculus: A Theory of Deterministic Queueing Systems for the Internet.* Springer-Verlag, 2001.
15. K. Lee. Performance bounds in communication networks with variable-rate links. In *Proc. ACM SIGCOMM'95*, pages 126–136, 1995.
16. C. Li, A. Burchard, and J. Liebeherr. A network calculus with effective bandwidth. Technical report, CS-2003-20, University of Virginia, November 2003.
17. J. Liebeherr. IWQoS 2004 Panel Talk: Post-Internet QoS Research, 2004.
18. Y. Liu, C.-K. Tham, and Y. Jiang. A stochastic network calculus. Technical report, ECE-CCN-0301, National University of Singapore, December 2003.
19. Y. Liu, C.-K. Tham, and Y. Jiang. Conformance study for networks with service level agreements. *Computer Networks*, 2005. in press.
20. R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Proc. of the Cambridge Philosophical Society*, 58(3):497–520, 1962.
21. D. Starobinski and M. Sidi. Stochastically bounded burstiness for communication networks. *IEEE Trans. Information Theory*, 46(1):206–212, Jan. 2000.
22. Workshop Attendees. Report of the National Science Foundation Workshop on Fundamental Research in Networking, April 2003.
23. Q. Yaron and M. Sidi. Performance and stability of communication networks via robust exponential bounds. *IEEE/ACM Trans. Networking*, 1:372–385, June 1993.
24. Q. Yin, Y. Jiang, S. Jiang, and P. Y. Kong. Analysis on generalized stochastically bounded bursty traffic for communication networks. In *Proc. IEEE LCN'02*, 2002.