Stochastic Service Curve and Delay Bound Analysis: A Single Node Case

Yuming Jiang
Norwegian University of Science and Technology (NTNU), Norway

***Abstract*—A packet-switched network node with constant capacity (in bps) is considered, where packets within each flow are served in the first in first out (FIFO) manner. While this single node system is perhaps the simplest computer communication system, its stochastic service curve characterization and independent case analysis in the context of stochastic network calculus (snetcal) are still basic and many crucial questions surprisingly remain open. Specifically, when the input is a single flow, what stochastic service curve and delay bound does the node provide? When the considered flow shares the node with another flow, what stochastic service curve and delay bound does the node provide to the considered flow, and if the two flows are independent, can this independence be made use of and how? The aim of this paper is to provide answers to these fundamental questions.**

I. Introduction

Network calculus is a theory dealing with queueing type problems encountered in packet-switched computer networks. To simplify the analysis, an important idea in network calculus is to characterize the traffic and service processes using some bounds and perform analysis based on such bounds. Network calculus has developed along two tracks — deterministic and stochastic. Deterministic network calculus, coined by [10], has been extensively studied since its introduction in early 1990s, and is nicely covered by two books [5] [27]. Stochastic network calculus is the probabilistic extension or generalization of deterministic network calculus. The development of *stochastic network calculus* (SNC) began also in early 1990s. Early representative works include [26][33][4] for traffic modeling, and [28] for server modeling. The book [5] also covers the theory of effective bandwidth, a first approach to SNC. However, due to challenges specific to stochastic networks, it is recently that crucial network calculus properties have been proved for SNC, e.g. [3][8][29][16][31][22][12]. A selection of recent results can be found in the book [21]. In addition, three surveys/overviews are available [32][13][20].

In SNC, stochastic service curve is the fundamental concept for server modeling. If some flows and servers are independent, it is expected that tighter analytical bounds can be obtained by making use of this independence information in the analysis. In this paper, we consider a work-conserving constant capacity node serving flows. Each flow consists of a sequence of packets that are served in the first in first out (FIFO) manner. This single node system is perhaps the simplest computer communication system. For such a system, an immediate impression is perhaps that it has been thoroughly investigated and is well understood, given the current snetcal literature. Unfortunately, this impression has no solid supporting ground and can hence be highly misleading. Indeed, while the snetcal literature has a lot of results based on its various stochastic arrival curve and stochastic service curve models, the following questions remain largely open. What stochastic service curve and delay bound does the node provide when the input is a single flow? What stochastic service curve and delay bound does the node provide when the considered flow shares the node with another flow? If the two flows are independent, can this independence be made use of and how?

The objective of this paper is to derive results providing answers to these fundamental questions. Specifically, (i) when there is only one flow, we prove a stochastic service curve (SSC) that has a bounding function equal to the complimentary cumulative distribution function (CCDF) of the packet length distribution. In addition to the delay bound directly obtained from the existing snetcal results and this SSC, an improved delay bound is derived, which is consistent with a result in the deterministic network calculus literature: *For delay bound analysis, the last packetizer may be ignored*[5][27]. (ii) When there is cross traffic, i.e., the node is shared by the traversing flow with a crossing flow, we prove that the node provides to the aggregate of the two flows an *aggregate behavior* stochastic service curve that also has a bounding function equal to the CCDF of the packet length distribution of the aggregate. Based on this and existing snetcal results, an SSC for the traversing flow is found. To overcome the potential difficulty in finding the packet length CCDF of the aggregate flow, a new and improved SSC is derived, where the flow independence can also be made use of. Moreover, in addition to delay bounds from these SSCs, an improved delay bound is obtained, where the flow independence information can be exploited. (iii) To illustrate the obtained delay bounds, two examples are provided. For the single flow case, the (best) bound matches with the exact result for $M/M/1$/FIFO. For the case with cross traffic, the obtained (best) bound is close to the exact result for $M/M/1$/priority.

The rest is structured as follows. In the next section, the system model and notation are defined. In Sec. III, stochastic network calculus basics are given. In Sec. IV, the difficulties for stochastic service curve and delay bound analysis, when packetization effect is not ignored, are discussed. In Sec. V, we focus on the single flow case. In Sec. VI, we take cross traffic into consideration, and find stochastic service curves and delay bounds for the traversing flow. In Sec. VII, we give examples. In Sec. VIII, discussion on related work is provided. Finally, concluding remarks are given in Sec. IX.

II. The System Model and Notation

We consider a work-conserving network node serving flows in a packet-switched network. It is a discrete-time system with time indexed by $t = 0, 1, 2, \ldots$. The serving capacity of the node is constant, denoted by C (in bps). Flow f traverses this node and is referred as the *traversing flow*. In addition, the node may also serve another flow f^c, which is the aggregate flow of crossing traffic and is referred as the *crossing flow*. Packets within each flow are served in the FIFO manner. Between flows, some scheduling policy is employed, but within this paper, it is not specified.

By convention, a packet is said to have arrived to (respectively served by) the node when and only when its last bit has arrived to (respectively left) the node. When a packet arrives seeing the node busy, the packet will be queued and the buffer size for such a queue is assumed to be large enough ensuring no packet loss. All queues are initially empty.

For the traversing flow f, we let $p^{f,i}$ denote the ith packet $(i = 1, 2, \ldots)$ of the flow. For each $p^{f,i}$, we denote by $a^{f,i}$ its arrival time to the node, $d^{f,i}$ its departure time from the node, and $l^{f,i}$ its length (in bits). Similarly, for the crossing flow f^c, we let $p^{c,i}$ denote its ith packet $(i = 1, 2, \ldots)$, $a^{c,i}$ its arrival time, $d^{c,i}$ its departure time, and $l^{c,i}$ its length.

We further use $A^f(t)$ and $A^c(t)$ to denote the amount of traffic (in bits) that has arrived from the traversing flow and the crossing flow to the node within time period $[0, t]$ respectively. Correspondingly, $A^f(s, t) = A^f(t) - A^f(s)$ and $A^c(s, t) = A^c(t) - A^c(s)$ respectively denote the amount of traffic (in bits) that has arrived from them within time period $(s, t]$. For the departures from the node, we use $A^{*f}(t)$ and $A^{*c}(t)$ to respectively denote the amount of traffic (in bits) that has been served from the traversing flow and the crossing flow within time period $[0, t]$. Correspondingly, $A^{*f}(s, t) = A^{*f}(t) - A^{*f}(s)$ and $A^{*c}(s, t) = A^{*c}(t) - A^{*c}(s)$ respectively represent the amount of traffic (in bits) that has been served from the traversing flow and the crossing flow by the node within time period $(s, t]$.

For the node, if it is shared by f and f^c, consider *the sequence of packets on the output link*. For this sequence of packets, we call it the *aggregate flow* at the node, and let $p^{g,j}$ denote the jth packet $(j = 1, 2, \ldots)$ of the aggregate flow. For each $p^{g,j}$, denote by $a^{g,j}$ the arrival time of the corresponding packet to the node, $d^{g,j}$ its departure time from the node, and $l^{g,j}$ its length (in bits). Note that the aggregate flow is resulted from the aggregation of the traversing flow and the crossing flow through the work-conserving constant capacity node. In addition, for the departures from the node, we use $A^{*g}(t)$ to denote the amount of traffic (in bits) from the aggregate flow, which has been served by the node within time period $[0, t]$, and $A^{*g}(s, t) = A^{*g}(t) - A^{*g}(s)$ the amount of traffic (in bits) that has been served from the aggregate flow by the node within time period $(s, t]$. It is worth highlighting that for the departures, there holds $A^{*g}(s, t) = A^{*f}(s, t) + A^{*c}(s, t)$.

The delay of $p^{f,i}$, denoted by $D^{f,i}$, is naturally

$$D^{f,i} \quad = \quad d^{f,i} - a^{f,i}. \tag{1}$$

In addition, we define the (virtual) delay at time t as

$$D^f(t) \quad = \quad \inf\{\tau : A^{*f}(t + \tau) \geq A^f(t)\}. \tag{2}$$

Due to FIFO, the delay of $p^{f,i}$ is also found from[1]

$$D^{f,i} = D^f(a^{f,i}) = \inf\{\tau : A^{*f}(a^{f,i} + \tau) \geq A^f(a^{f,i})\}. \tag{3}$$

The min-plus convolution, denoted by $\otimes$, of functions $f(\cdot)$ and $g(\cdot)$, is defined as:

$$f \otimes g(y) = \inf_{0 \leq x \leq y} \{f(x) + g(y - x)\} \tag{4}$$

and it is easily verified $f \otimes g(y) = g \otimes f(y)$.

The maximum horizontal distance between functions $\alpha(\cdot)$ and $\beta(\cdot)$, denoted by $h(\alpha, \beta)$, is defined as

$$h(\alpha, \beta) = \sup_{s \geq 0} \{\inf\{\tau \geq 0 : \alpha(s) \leq \beta(s + \tau)\}\}. \tag{5}$$

III. Stochastic Network Calculus Basics

In this section, some related stochastic network calculus models and existing results are introduced.

A. Models

In snetcal, stochastic arrival curve (SAC) and stochastic service curve (SSC) are the most fundamental models. While SAC is for traffic modeling, SSC is for server modeling. In the literature, there are several definition variations of SAC and SSC. In this paper, we adopt the following, to which the other variations may be mapped [21].

Definition 1. *A flow is said to have a v.b.c (virtual backlog centric) stochastic arrival curve $\alpha(t)$ with bounding function $\bar{F}$, if its arrival process $A(t)$ satisfies, for any $t \geq 0$ [11][34][22],*

$$P\{A(t) - A \otimes \alpha(t) > x\} \leq \bar{F}(x) \tag{6}$$

where $\alpha(t)$ is non-negative non-decreasing on t, and $\bar{F}(x)$ non-negative non-increasing on x.

In Definition 1, if $\bar{F}(0) = 0$, implying $\bar{F}(x) = 0$ for all $x \geq 0$ or in other words $A(t) \leq A \otimes \alpha(t)$, $\alpha(t)$ is also called a (deterministic) arrival curve of the flow in the network calculus literature.

Definition 2. *A system is said to provide a stochastic service curve $\beta(t)$ with bounding function $\bar{G}$, if there holds, for all $t \geq 0$ [11] [21],*

$$P\{A \otimes \beta(t) - A^*(t) > x\} \leq \bar{G}(x) \tag{7}$$

where $\beta(t)$ is non-negative non-decreasing on t, and $\bar{G}(x)$ non-negative non-increasing on x.

In Definition 2, if $\bar{G}(0) = 0$, implying $\bar{G}(x) = 0$ for all $x \geq 0$ or in other words $A^*(t) \geq A \otimes \beta(t)$, $\beta(t)$ is also called a (deterministic) service curve of the system in the network calculus literature.

[1] Strictly speaking, $D^{f,i} \leq D^f(a^{f,i})$, where the equation holds only if there is no concurrent arrival at $a^{f,i}$. If $A(t)$ and $A^*(t)$ are defined on $[0, t)$, this virtual delay definition defines the virtual waiting time for a (possible virtual) arrival at time t.

B. Related Results

This paper focuses on stochastic service curve and delay bound analysis. The following presents some related results.

For stochastic service curve analysis, due to the difficulties that will be discussed in the next section, very little is known for the general case where packet length distribution is taken into consideration, and available results mostly assume fluid system ignoring packetization effect or that all packets have the same length.

In the SNC literature, the following result, called the *leftover service* property, has been widely used for finding the stochastic service curve charaterization of the service provided to a flow (e.g. see [8][31]).

Proposition 1. *Consider a system with cross traffic. If the system provides a stochastic service curve $\beta(t)$ with bounding function $\bar{G}(x)$ and the crossing flow has a v.b.c. stochastic arrival curve $\alpha^c(t)$ with bounding function $\bar{F}^c(x)$, then the* leftover service *provided to the traversing flow has a stochastic service curve $\beta^f(t) = (\beta(t) - \alpha^c(t))^+$ with bounding function $\bar{G}^f(x) = \bar{F}^c \otimes \bar{G}(x)$.*

For delay bound analysis, the following result has been proved (e.g. see [11][21]).

Proposition 2. *If a system provides a stochastic service curve $\beta(t)$ with bounding function $\bar{G}$ to a flow f, which has v.b.c stochastic arrival curve $\alpha^f(t)$ with bounding function $\bar{F}^f$, then the flow has a delay bound as*

$$P\{D^f(t) > h(\alpha^f + x, \beta)\} \le \bar{F}^f \otimes \bar{G}(x). \tag{8}$$

When the system is shared by the traversing flow and the crossing flow, the following delay bound follows immediately from Proposition 1 and Proposition 2 [21].

Proposition 3. *Under the same condition as Preposition 1, if the traversing flow has a stochastic arrival curve $\alpha^f(t)$ with bounding function $\bar{F}^f(x)$, then the delay of the flow is bounded as*

$$P\{D^f(t) > h(\alpha^f + x, \beta^f)\} \le \bar{F}^f \otimes \bar{F}^c \otimes \bar{G}(x). \tag{9}$$

where $\beta^f(t) = (\beta(t) - \alpha^c(t))^+$.

IV. The Difficulties

As highlighted in the previous section, stochastic service curve (SSC) is the most fundamental server model for snetcal. As reviewed there, if the SSC characterization of the service provided by the node to the flow and the v.b.c SAC characterization of the flow are known, a delay bound can be readily obtained from the existing snetcal results. In the literature, while many results (e.g. [19][22][23]) may be exploited to find the v.b.c SAC characterization of a flow, there are very few for SSC analysis.

The difficulties are inherent in the SSC definition. Suppose $S(t)$ is the service process provided by the node to the flow. The following equation, called the min-plus convolution queueing principle [18], holds [21]:

$$A^*(t) = \inf_{0 \le s \le t}\{A(s) + S(s,t)\} \tag{10}$$

where $S(t)$ denotes the service process provided to the flow and $S(s,t) \equiv S(t) - S(s)$.

Essentially, SSC defines a way to characterize the service process $S(t)$. While the SSC definition allows the derivation of results useful for performance study of computer networks, finding the SSC characterization of a system is surprisingly challenging. Even for the simplest constant capacity single node system, the challenge already exists.

A. A Pitfall

When the node has constant capacity C (in bps), the following equation has sometimes been *wrongly* believed in the network calculus literature (see, e.g., [25]),

$$A^*(t) = \inf_{0 \le s \le t}\{A(s) + C \cdot (t - s)\}. \tag{11}$$

Or in other words, it is *wrongly* believed, for the constant capacity node:

$$S(s,t) = C \cdot (t - s). \tag{12}$$

To give a counterexample, let's consider a single packet flow input. The packet arrives at time $a^{f,1} = 1$ and has length 2. It is then clear that $A(0) = 0, A(1) = 2, A(2) = 2, A(3) = 2$. Suppose $C = 1$. Then, $d^{f,1} = 3$. Hence, $A^*(0) = 0, A^*(1) = 0, A^*(2) = 0, A^*(3) = 2$. However, from (11), the output would be $A^*(0) = 0, A^*(1) = 1, A^*(2) = 2, A^*(3) = 2$, which is wrong[2]. Table I summarizes the comparison.

TABLE I
A COUNTEREXAMPLE

t	0	1	2	3	4
$A(t)$	0	2	2	2	2
$A^*(t)$ actual	0	0	0	2	2
$A^*(t)$ from (11)	0	1	2	2	2

B. Difficulty in Finding SSC

When fluid-flow is assumed, i.e. $l^{f,i} \to 0$ and $l^{c,j} \to 0$ for all packets, it is easy to verify that the amount of service provided by the node during any backlog period with length τ is $C \cdot \tau$. Then, from the network calculus literature, it is known that the node provides a deterministic service curve $\beta(t) = C \cdot t$. In addition, when cross traffic is present, the stochastic service curve provided to the traversing flow is readily derived from the network calculus *leftover service* property as shown by Proposition 1. In addition, from the traversing flow's viewpoint, the crossing flow can be treated as a process that impairs the total service provided by the server. Then with the impairment process concept [16][21], the stochastic service curve characterization of the service available to the traversing flow can also be found.

[2] Choosing to define $A(t)$ and $A^*(t)$ on $[0, t)$ does not correct the mistake.

However, when packetization effect is taken into account, finding stochastic service curves for the node becomes *surprisingly* challenging, even though it has constant capacity.

Indeed, the network calculus literature has shown that a constant server with capacity C has a deterministic service curve $(C \cdot t - L^{max})^+$, where L^{max} denotes the maximum packet length in the system. This follows from two fundamental results. (i) If there is a function that lower-bounds the amount of service provided to the input during any backlog period, then the function is a (deterministic) service curve of the server [27]. (ii) Within any backlog period of length t, the amount of service provided by a constant rate server with capacity C is lower bounded by $(C \cdot t - L^{max})^+$ [14].

Fundamentally, the following inequality can be proved [14]:

$$A^*(t) \geq A(t) \otimes (C \cdot t - L^{max}(t))^+. \qquad (13)$$

where $L^{max}(t) \equiv \max\{l^1, l^2, l^{(t)}\}$ with $l^{(t)}$ denoting the length of the most recent packet that arrived before or at t, and $L^{max} = \lim_{t\to\infty} L^{max}(t)$.

With simple manipulation based on the definition of $\otimes$, we obtain

$$A(t) \otimes (C \cdot t) - A^*(t) \leq L^{max}(t) \qquad (14)$$

which implies

$$\begin{aligned} P\{A(t) \otimes (C \cdot t) - A^*(t) > x\} &\leq P\{L^{max}(t) > x\} \\ &\equiv \bar{F}^{L^{max}(t)}(x). \qquad (15) \end{aligned}$$

Then, we can conclude that *the constant capacity node provides a stochastic service curve $C \cdot t$ with bounding function $\bar{F}^{L^{max}(t)}(x)$*.

Unfortunately, $L^{max}(t)$ is non-decreasing with t, implying that $\bar{F}^{L^{max}(t)}(x)$ may approach 1 as t grows [9][3]. Consequently, using $\bar{F}^{L^{max}(t)}$ as a bounding function is meaningless.

The problem becomes even more challenging when there is cross traffic. First, in order to apply the leftover service property to obtain a stochastic service curve for the traversing flow, we need to know the stochastic service curve of the node. However, the above discussion implies that the stochastic service curve of the node is yet to be found. Second, the packet length process is a mixture of the packet length process of the traversing flow and that of the crossing flow. This makes the determination of $\bar{F}^{L^{max}(t)}$ and consequently the stochastic service curve characterization of the node even more difficult.

To tackle the time-growing $\bar{F}^{L^{max}(t)}$ problem, one may introduce a compromised service curve $(C - \theta) \cdot t$, for some $\theta \geq 0$, with a resultant bounding function related to $\int_x^\infty \bar{F}^l(y)dy$, by exploiting an approach used in SNC in dealing with maximal random processes [21]. Recently, the effort in [30] shows that, without compromising the service curve expression $C \cdot t$, a bounding function, which is also related to

$$\int_x^\infty \bar{F}^l(y)dy$$

can be found, when the packet length process is stationary and satisfies some conditions.

Note that intuitively, a packetized system can be treated as the concatenation of a fluid system followed by a packetizer [5][27] [4]. Since the fluid system provides deterministic service curve $C \cdot t$ as discussed above, the stochastic behavior of the node is hence determined by the packet length distribution. Based on this, we boldly conjecture that *the constant capacity node provides a stochastic service curve $C \cdot t$ with bounding function simply as $\bar{F}^l$*. However, while the intuition is perhaps straightforward, proving the validity of the conjecture is far from direct as to be shown in the next section.

C. Difficulty in Making Use of Independence Information

Besides the difficulty in finding the SSC characterization of the node, it is even more difficult to make use of potential independence information in the analysis. This is due to that, the service process $S(t)$ and the arrival process $A(t)$ are inherently dependent, implied by (10). More specifically, both $S(t)$ and $A(t)$ are functions of the lengths of packets that are counted in. For a simple example, suppose flow f only has one packet $p^{f,1}$ whose length $l^{f,1}$ is a random variable. It is clear that $A(t) = l^{f,1}$ and also $S(t) = l^{f,1}$ for any $t \geq d^{f,1}$, which indicates strong dependence between $A(t)$ and $S(t)$.

The inherent dependence between $A(t)$ and $S(t)$ makes it difficult to make use of potential independence in the analysis. Particularly, when there is cross traffic present, even though the traversing flow may be independent of the crossing flow, this independence information cannot be exploited when applying the existing snetcal results as reviewed in Section III. This is due to that, the stochastic service curve characterization of the service process $S(t)$ provided by the node is dependent on the packet lengths of both flows. A consequence is that, with the SSC decided from Proposition 1, it is not possible to make use of the dependence information to improve the (independence-information-unaware) delay bound in Proposition 3.

To this point, we would like to remark that, if all packets have the same length[5] or their lengths are upper-bounded, the service process of the node has a deterministic service curve. For this case, independent case analysis may be conducted by following the approaches proposed in [16] and [12].

However, for the more general case, even though it is intuitive that the independence information of the two flows should allow improving the analysis, how specifically to make use of this independence information in the analysis remains to be addressed.

[3] An exception is when all packets have the same length L or their lengths are upper-bounded by L. In this case, for any t, $\bar{F}^{L^{max}(t)}(x) \leq 1$ for all $x \leq L$; otherwise $\bar{F}^{L^{max}(t)}(x) = 0$ for all $x > L$. Under this case, the node provides a deterministic service curve $(C \cdot t - L)^+$, with which, further analysis similar to that under the fluid-flow case can be conducted.

[4] A packetizer is an element gathering all bits in a packet, which delivers the entire packet with no delay until and immediately after receiving the last bit of the packet.

[5] The fluid-flow is a special case with infinitely small packet length.

V. Stochastic Service Curve and Delay Bounds: The Single Flow Case

In this section, we first prove the stochastic service curve as suggested by the conjecture. Then, delay bounds are derived for the traversing flow. To deal with the difficulty in finding the SSC, a novel approach is introduced.

A. Stochastic Service Curve

We now present the approach to tackling the difficulty. In this approach, we relate the service provided by a (not-necessarily constant rate) system to a flow f to a virtual time function defined as

$$V^{f,i}(R) \quad = \quad \max\{a^{f,i}, V^{f,i-1}\} + \frac{l^{f,i}}{R} \qquad (16)$$

iteratively for $i = 1, 2 \ldots,$ with $V^{f,0} = 0$, where R is a constant rate parameter.

Applying iteratively to its right hand side, (16) becomes

$$V^{f,i}(R) \quad = \quad \max_{1 \le j \le i}\{a^{f,j} + \frac{\sum_{k=j}^{i} l^{f,k}}{R}\}. \qquad (17)$$

The following result is crucial, which establishes a link between the stochastic service curve model and the virtual time function. For deterministic network calculus, a similar relationship can be found in [15] (Lemma 2). The proof is long and is included in the appendix.

Lemma 1. *Consider a flow f served by a system. For any time $t > 0$ and $R > 0$, the following relationship holds (for any sample path of the system):*

$$A^f \otimes \beta(t) - A^{f*}(t) \le R \cdot [d^{f,i(t)} - V^{f,i(t)}(R)] + l^{f,i(t)} \quad (18)$$

where $\beta(t) = R \cdot t$, $i(t) = \min\{k : d^{f,k} \ge t\}$[6]*, and $l^{f,i(t)}$ the length of packet $p^{f,i(t)}$.*

For the considered single node system with single input flow f, consider any packet $p^{f,i}$. There are two cases. One case is that when $p^{f,i}$ arrives, the system is idle, which is $a^{f,i} > d^{f,i-1}$. Hence, $d^{f,i} = a^{f,i} + \frac{l^{f,i}}{C}$. Another case is that $p^{f,i}$ arrives, the system is busy, which is $a^{f,i} \le d^{f,i-1}$. Then, it has to wait until the previous packet $p^{f,i-1}$ has finished service. Hence, $d^{f,i} = d^{f,i-1} + \frac{l^{f,i}}{C}$. Combining both cases, we must have $d^{f,i} = \max\{a^{f,i}, d^{f,i-1}\} + \frac{l^{f,i}}{C}$. Comparing it with $V^{f,i}(C)$, the following result is proved.

Lemma 2. *For the considered single node system with single input flow f, there holds, for any packet $p^{f,i}$ of the flow,*

$$d^{f,i} = V^{f,i}(C). \qquad (19)$$

Applying (19) to Lemma 1, the following is obtained for $R = C$:

$$A^f \otimes \beta(t) - A^{f*}(t) \le l^{f,i(t)}. \qquad (20)$$

It is worth highlighting that $i(t)$ is random and packet $p^{f,i(t)}$ may be different from one sample path to another sample path. However, if all packets, $l^{f,1}, l^{f,2}, \ldots,$ have identically distributed packet lengths with CCDF $\bar{F}^l(x)$, or more generally if their lengths have the same upper-bounded CCDF $\bar{F}^l(x)$, the following follows from (20).

$$P\{A^f \otimes \beta(t) - A^{f*}(t) > x\} \le \bar{F}^l(x). \qquad (21)$$

Summarizing the above discussion, we have validated the conjecture. Formally, the following theorem has been proved:

Theorem 1. *Consider a work-conserving system with constant capacity C serving a flow f. Suppose that all packets have length distributions that are identical with CCDF $\bar{F}^l(x)$ or whose CCDFs are all upper-bounded by $\bar{F}^l(x)$. Then, the system provides to the flow a stochastic service curve $\beta(t) = C \cdot t$ with bounding function $\bar{G}(x) = \bar{F}^l(x)$.*

B. Delay Bounds

With Theorem 1, the following delay bound follows directly from Proposition 2.

Corollary 1. *Under the same condition as for Theorem 1, if the traversing flow has a v.b.c stochastic arrive curve $\alpha(t) = r^f \cdot t$ with bounding function $\bar{F}^f$ and $r^f \le C$, then for any packet $p^{f,i}$, its delay is bounded as:*

$$P\{D^{f,i} > \tau\} \le \bar{F}^l \otimes \bar{F}^f(C \cdot \tau) \qquad (22)$$

In Section IV, we have discussed the inherent dependence between the arrival process and the service process. When it comes to the delay bound analysis, we would like to highlight that the inherent dependence is specifically seen between $A^f(t)$ and $l^{f,i(t)}$, since by their definitions, $l^{f,i(t)}$ may be counted in $A^f(t)$. This partly explains why the min-plus convolution appears on the right hand side of (22) and in Proposition 2, which assumes no knowledge of potential independence information.

At this moment, it seems that nothing more than Corollary 1 could be done for delay bound analysis. In the following, we show that this is too pessimistic. Specifically, by exploiting the idea of the virtual time function, an improved delay bound is proved in the following theorem.

Theorem 2. *Under the same condition as for Theorem 1, if the traversing flow has a v.b.c stochastic arrive curve $\alpha(t) = r^f \cdot t$ with bounding function $\bar{F}^f$ and $r^f \le C$, then for any packet $p^{f,i}$, its delay is bounded as (a.s.):*

$$P\{D^{f,i} > \tau\} \le \bar{F}^f(C \cdot \tau). \qquad (23)$$

Proof: Consider any sample path of the system. By the definition of $D^{f,i}$ and with Lemma 2, we have

$$\begin{aligned} & D^{f,i} = d^{f,i} - a^{f,i} = V^{f,i}(C) - a^{f,i} \\ = \; & \frac{1}{C} \max_{1 \le j \le i}\{\sum_{k=j}^{i} l^{f,k} - C(a^{f,i} - a^{f,j})\} \\ \le \; & \frac{1}{C} \max_{0 \le j \le i}\{A^f(a^{f,j} - \epsilon, a^{f,i}) - C(a^{f,i} - a^{f,j} + \epsilon)\} + \epsilon \\ & \qquad (24) \\ \le \; & \frac{1}{C} \sup_{0 \le s \le t}\{A^f(s,t) - r^f(t-s)\} + \epsilon \qquad (25) \end{aligned}$$

[6] Intuitively, if at time t, there is a packet under service from the flow, then $p^{f,i(t)}$ is this packet; otherwise, $p^{f,i(t)}$ is the first packet from this flow, which receives service after t.

where $t = a^{f,i}$, $\epsilon \to 0$ and $r^f \le C$. In step (24), $\epsilon \to 0$ is introduced such that $A^f(a^{f,j} - \epsilon, a^{f,i})$ includes all arrivals in $[a^{f,j}, a^{f,i}]$.

Since the traversing flow has a v.b.c stochastic arrive curve, we have by definition:

$$P\{\sup_{0 \le s \le t} \{A^f(s,t) - r^f(t-s)\} > x\} \le \bar{F}^f.$$

Since this bounding function holds for all sample paths, (23) is then obtained. ■

At a first glance, the delay bound in Theorem 2 may seem to be surprising, since the packetization effect is not directly seen from (23). However, an alert reader may have noticed that it is indeed consistent with a result in the deterministic network calculus literature, which states that in delay bound analysis, *the last packetizer on the path of the flow may be ignored* [5][27]. Theorem 2 proves this property in the context of stochastic network calculus for the single node case.

Remark: An implication of Theorem 2 is that, when delay bound analysis is performed, the node may be treated as if it would provide a deterministic service curve $C \cdot t$ and then Corollary 1 becomes the same as Theorem 2.

VI. Stochastic Service Curves and Delay Bounds: The Case with Cross Traffic

In this section, we consider the case where the traversing flow shares service of the node with the crossing flow. Specifically, we find a stochastic service curve for the node and two SSCs for the traversing flow, followed by deriving delay bounds for the traversing flow.

A. Stochastic Service Curves

1) A direct result: Let us treat the traversing flow and the crossing flow as an aggregate flow. For packets of the aggregate flow g, *which takes the packet order as that on the output link,* the following relation can be easily verified:

$$d^{g,j} = \max\{a^{g,j}, d^{g,j-1}\} + \frac{l^{g,j}}{C}. \qquad (26)$$

Comparing (26) with (16), it is clear that for the aggregate, $d^{g,j} = V^{g,j}(C)$.

Note that in presenting (26), we do not make any assumption on the scheduling algorithm between the two flows, and (26) is only concerned about the aggregate. We call (26) the "aggregate behavior" of the node, which is in consistence with the definition of the aggregate per-hop behavior [17] under the Differentiated Services (DiffServ) architecture [1].

With (26) and following the same proof of Theorem 1, the following result can be verified.

Lemma 3. *Consider a work-conserving system with constant capacity C, shared by a traversing flow f and a crossing flow f^c. Suppose that all packets have length distributions that are identical with CCDF $\bar{F}^{l^g}(x)$ or whose CCDFs are all upper-bounded by $\bar{F}^{l^g}(x)$. Then, the system provides to the aggregate of the two flows an "aggregate behavior" stochastic service curve $C \cdot t$ with bounding function $\bar{F}^{l^g}(x)$.*

Recall that we are interested in the traversing flow. With Lemma 3 and the leftover property Proposition 1, the following stochastic service curve to the traversing flow is obtained[7]:

Theorem 3. *Consider the same system as in Lemma 3. If the the crossing flow has a v.b.c. stochastic arrival curve $\alpha^c(t) = r^c \cdot t$, ($r^c < C$), with bounding function $\bar{F}^c(x)$, then the system provides to the traversing flow a stochastic service curve $\beta(t) = (C - r^c) \cdot t$ with bounding function $\bar{G}(x)$ as*

$$\bar{G}(x) \quad = \quad \bar{F}^c \otimes \bar{F}^{l^g}(x) \qquad (27)$$

Note that in Theorem 3, the resulting bounding function is $\bar{F}^c \otimes \bar{F}^{l^g}(x)$, which assumes no knowledge of potential independence information, even though the crossing flow may be independent of the traversing flow. This is due to that F^{l^g} is the length distribution of all packets, which include packets of the crossing flow, and hence F^{l^g} is inherently coupled with the traffic arrival process of the crossing flow.

2) An improved result: While Theorem 3 is an improvement over those that are based on (15), it may be difficult to find $\bar{F}^{l^g}$ of the aggregate particularly when the traversing flow and the crossing flow have different packet length distributions.

The following theorem proves another stochastic service curve for the traversing flow, where there is no need to find $\bar{F}^{l^g}$, relieving the difficulty. In addition, if the two flows are independent, this independence information is made use of.

Theorem 4. *Consider a work-conserving system with constant capacity C, shared by a traversing flow f and a crossing flow f^c. Suppose the crossing flow has a v.b.c stochastic arrival curve $\alpha^c(t) = r^c \cdot t$, ($r^c < C$), with bounding function $\bar{F}^c$, and suppose all packets of the traversing flow have length distributions that are identical with CCDF $\bar{F}^l(x)$ or whose CCDFs are all upper-bounded by $\bar{F}^l(x)$. Then, the node provides to the traversing flow a stochastic service curve $\beta(t) = (C - r^c)$ with bounding function $\bar{G}(x)$ as*

$$\bar{G}(x) \quad = \quad \bar{F}^c \otimes \bar{F}^l(x) \qquad (28)$$

and if the two flows are independent,

$$\bar{G}(x) \quad = \quad 1 - F^c * F^l(x) \qquad (29)$$

where, $F^l \equiv 1 - \bar{F}^l$, $F^c \equiv 1 - \bar{F}^c$, and $F_1 * F_2(x) \equiv \int_0^x F_1(x - y)dF_2(y)$.

To prove Theorem 4, Lemma 4 and Lemma 5 below are crucial, with which, Theorem 4 is easily verified.

Lemma 4. *For the considered single node system with cross traffic, there holds, for any packet $p^{f,i}$ of the traversing flow,*

$$d^{f,i} \le V^{f,i}(C - r^c) + \frac{\sup_{0 \le s \le d^{f,i}}\{A^c(s, d^{f,i}) - r^c(d^{f,i} - s)\}}{C - r^c} \qquad (30)$$

for any $C > r^c \ge 0$.

Proof: As for (26), let us consider the aggregate flow g. Since no specific scheduling between the two flows has been

[7] Strictly speaking, instead of directly applying Proposition 1, a separate proof is needed.

assumed, a packet, which appears earlier on the output link in the aggregate flow, may actually arrive to the node later than another packet that appears later on the output link. In other words, we may not have $a^{g,j} \geq a^{g,j-1}$.

For any packet $p^{f,i}$, suppose its corresponding packet in the aggregate flow g is $p^{g,j}$. Particularly, we suppose the departure time of $p^{f,i}$, i.e. $d^{f,i} = d^{g,j}$, is within the busy period that starts at t^0. Note that such a busy period always exists, since in the worst case, the period is only the service time period of $p^{f,i}$ and in this case, $t^0 = a^{g,j}$.

Since the node is work-conserving with constant service rate C and it is busy with serving between t^0 and $d^{g,j}$, there holds:

$$d^{g,j} = t^0 + \frac{\sum_{k=j_0}^{j} l^{g,k}}{C}, \tag{31}$$

where p^{g,j_0} denotes the packet whose arrival starts the busy period.

Among packets $p^{g,j_0}, \ldots, p^{g,j}$, some belong to the traversing flow and the rest the crossing flow. Let p^{f,i_0} denote the first packet from the traversing flow served in the busy period. There holds $a^{f,i_0} \geq t^0$.

Equation (31) can be re-written as:

$$d^{f,i} \leq t^0 + \frac{\sum_{k=i_0}^{i} l^{f,k}}{C} + \frac{A^{*c}(t^0, d^{f,i})}{C}, \tag{32}$$

where, by definition, $A^{*c}(t^0, d^{f,i})$ represents the total length (in bits) of packets from the crossing flow served in $(t^0, d^{f,i}]$[8].

Since the busy period starts at t^0, this implies that immediately before t^0, the node is idle. In other words, all packets, which arrived before t^0, have been served by t^0. So, we have $A^{*c}(t^0) = A^c(t^0)$. In addition, crossing flow packets, which are served before $d^{f,i}$, must have arrived by $d^{f,i}$. So, we have $A^{*c}(d^{f,i}) \leq A^c(d^{f,i})$. Combing both, we obtain:

$$A^{*c}(t^0, d^{f,i}) \leq A^c(t^0, d^{f,i}) \tag{33}$$

which, when applied to (32), results in

$$d^{f,i} \leq t^0 + \frac{\sum_{k=i_0}^{i} l^{f,k}}{C} + \frac{A^{c}(t^0, d^{f,i})}{C}. \tag{34}$$

With (34), we obtain, for any $C > r^c \geq 0$,

$$\begin{aligned} d^{f,i} \quad &\leq \quad t^0 + \frac{\sum_{k=i_0}^{i} l^{f,k}}{C} + \frac{A^c(t^0, d^{f,i}) - r^c(d^{f,i} - t^0)}{C} \\ &\quad + \frac{r^c(d^{f,i} - t^0)}{C} \end{aligned}$$

Further with simple manipulation, we obtain

$$d^{f,i} \leq t^0 + \frac{\sum_{k=i_0}^{i} l^{f,k}}{C - r^c} + \frac{A^c(t^0, d^{f,i}) - r^c(d^{f,i} - t^0)}{C - r^c} \tag{35}$$

Recall the virtual time function (16), it is ease to verify that, for the considered packet $p^{f,i}$, we have

$$V^{f,i}(C - r^c) \quad \geq \quad a^{f,i_0} + \frac{\sum_{k=i_0}^{i} l^{f,k}}{C - r^c} \geq t_0 + \frac{\sum_{k=i_0}^{i} l^{f,k}}{C - r^c} \tag{36}$$

In addition, there holds

$$\begin{aligned} & A^c(t^0, d^{f,i}) - r^c(d^{f,i} - t^0) \\ \leq \quad & \sup_{0 \leq s \leq d^{f,i}} \{A^c(s, d^{f,i}) - r^c(d^{f,i} - s)\} \end{aligned} \tag{37}$$

Applying (36) and (37) to (35), we obtain (30) and the lemma is proved. ∎

We remark that when there is no cross traffic, i.e. $A^c(t) = 0$, then letting $r^c = 0$, Lemma 4 is reduced to Lemma 2 as expected.

Note that Lemma 1 provides a general relationship, with which, by letting $R = C - r^c$ in it, we obtain

$$\begin{aligned} & A^f \otimes \beta(t) - A^{f*}(t) \\ \leq \quad & (C - r^c)[d^{f,i(t)} - V^{f,i(t)}(C - r^c)] + l^{f,i(t)} \end{aligned} \tag{38}$$

Applying Lemma 4 to above immediately gives the following:

Lemma 5. *For the considered single node system with cross traffic, for any time t and any sample path of the system, the following relationship holds for the traversing flow f,*

$$\begin{aligned} & A^f \otimes \beta(t) - A^{f*}(t) \qquad (39) \\ \leq \quad & \sup_{0 \leq s \leq d^{f,i(t)}} \{A^c(s, d^{f,i(t)}) - r^c \cdot (d^{f,i(t)} - s)\} + l^{f,i(t)} \end{aligned}$$

where $\beta(t) = (C - r^c) \cdot t$, $i(t) = \min\{k : d^{f,k} \geq t\}$, and $l^{f,i(t)}$ is the length of packet $p^{f,i(t)}$.

Finally, since the crossing flow has a v.b.c stochastic arrival curve $r^c t$ with bounding function $\bar{F}^c$, and all packet lengths have identical (or the same upper-bounded) CCDF $\bar{F}_l$, Theorem 4 is proved by applying these conditions to Lemma 5.

It is worth highlighting that while (28) looks similar to (27), there is a fundamental difference between them. Specifically, the packet length distribution F^{l^g} in (27) is that of the aggregate flow, while in (28), the packet length distribution F^l is only of the traversing flow.

B. Delay Bounds

1) Delay bounds from Theorems 3 and 4: With Theorems 3 and 4, the following delay bounds are directly obtained from Preposition 2 respectively.

Corollary 2. *Under the same condition as for Theorem 3, if the traversing flow has a v.b.c stochastic arrival curve $\alpha(t) = r^f \cdot t$ with bounding function $\bar{F}^f$, and $r^f < C - r^c$, then for any packet $p^{f,i}$, it has a delay bound as:*

$$P\{D^{f,i} > \tau\} \leq \bar{F}^c \otimes \bar{F}^{l^g} \otimes \bar{F}^f((C - r^c)\tau). \tag{40}$$

Corollary 3. *Under the same condition as for Theorem 4, if the traversing flow has a v.b.c stochastic arrival curve $\alpha(t) = r^f \cdot t$ with bounding function $\bar{F}^f$, and $\rho^f < C - r^c$, then for any packet $p^{f,i}$, it has a delay bound as:*
(i) if the two flows may be dependent,

$$P\{D^{f,i} > \tau\} \leq \bar{F}^c \otimes \bar{F}^l \otimes \bar{F}^f((C - r^c)\tau); \tag{41}$$

[8] Note that t^0 starts the busy period and hence no packet finishes service at t^0. This implies $A^{*c}(t^0, d^{f,i})$ indeed represents the total length (in bits) of packets from the crossing flow served in $[t^0, d^{f,i}]$.

(ii) if the two flows are independent,

$$P\{D^{f,i} > \tau\} \leq (1 - F^c * F^l) \otimes \bar{F}^f((C - r^c)\tau). \qquad (42)$$

It is worth highlighting that in obtaining the bounding function in Theorem 4, we have relied on the right hand side of (39), which and $A(t)$ are inherently dependent due to $l^{f,i(t)}$. This explains why in (42), the independence information cannot be further made use of.

2) An improved delay bound: In the following, an improved delay bound is presented.

Theorem 5. *Suppose the traversing flow has a v.b.c stochastic arrival curve $\alpha(t) = r^f \cdot t$ with bounding function $\bar{F}^f$ and the crossing flow has a v.b.c stochastic arrival curve $\alpha(t) = r^c \cdot t$ with bounding function $\bar{F}^c$. If $r^f + r^c < C$, then for any packet $p^{f,i}$ of the traversing flow, its delay is bounded as (a.s.)*
(i) if the two flows may be dependent,

$$P\{D^{f,i} > \tau\} \leq \bar{F}^c \otimes \bar{F}^f((C - r^c)\tau); \qquad (43)$$

(ii) if the two flows are independent,

$$P\{D^{f,i} > \tau\} \leq 1 - F^c * F^f((C - r^c)\tau). \qquad (44)$$

Proof: Consider any sample path. With Lemma 4 particularly (30), we obtain, for any packet $p^{f,i}$,

$$\begin{aligned} & D^{f,i} = d^{f,i} - a^{f,i} \\ \leq\; & V^{f,i}(C - r^c) - a^{f,i} \\ & + \frac{\sup_{0 \leq s \leq d^{f,i}} \{A^c(s, d^{f,i}) - r^c \cdot (d^{f,i} - s)\}}{C - r^c} \\ =\; & \max_{1 \leq j \leq i} \{a^{f,j} + \frac{\sum_{k=j}^{i} l^{f,k}}{C - r^c}\} - a^{f,i} \\ & + \frac{\sup_{0 \leq s \leq d^{f,i}} \{A^c(s, d^{f,i}) - r^c \cdot (d^{f,i} - s)\}}{C - r^c} \qquad (45) \end{aligned}$$

To ease the presentation, we move $(C - r^c)$ to the left and get

$$\begin{aligned} & D^{f,i} \cdot (C - r^c) \\ \leq\; & \max_{1 \leq j \leq i} \{\sum_{k=j}^{i} l^{f,k} - (C - r^c)(a^{f,i} - a^{f,j})\} \\ & + \sup_{0 \leq s \leq d^{f,i}} \{A^c(s, d^{f,i}) - r^c \cdot (d^{f,i} - s)\} \\ \leq\; & \max_{0 \leq j \leq i} \{A^f(a^{f,j} - \epsilon, a^{f,i}) - (C - r^c)(a^{f,i} - a^{f,j} + \epsilon)\} \\ & + \sup_{0 \leq s \leq d^{f,i}} \{A^c(s, d^{f,i}) - r^c \cdot (d^{f,i} - s)\} + \epsilon \qquad (46) \\ \leq\; & \sup_{0 \leq s \leq a^{f,i}} \{A^f(s, a^{f,i}) - r^f(a^{f,i} - s)\} \\ & + \sup_{0 \leq s \leq d^{f,i}} \{A^c(s, d^{f,i}) - r^c \cdot (d^{f,i} - s)\} + \epsilon \qquad (47) \end{aligned}$$

where $\epsilon \to 0$.

Note that, given $a^{f,i}$ as implied by the delay definition, the first two terms on the right hand side of (47) are independent. This independence is more easily seen by expending them as

$$\begin{aligned} & \sup_{0 \leq s \leq a^{f,i}} \{A^f(s, a^{f,i}) - r^f(a^{f,i} - s)\} + \\ \max & \left\{ \begin{aligned} & \sup_{0 \leq s \leq a^{f,i}} \{A^c(s, a^{f,i}) - r^c \cdot (a^{f,i} - s)\} \\ & + A^c(a^{f,i}, d^{f,i}) - r^c \cdot (d^{f,i} - a^{f,i}), \\ & \sup_{a^{f,i} < s \leq d^{f,i}} \{A^c(s, d^{f,i}) - r^c \cdot (d^{f,i} - s)\} \end{aligned} \right\} \qquad (48) \end{aligned}$$

where the first term is determined only by the time period $[0, a^{f,i}]$ and the arrivals of the traversing flow in this period, while the second term is determined by the same period $[0, a^{f,i}]$ and another later non-overlapping period $(a^{f,i}, d^{f,i}]$ and the arrivals of the crossing flow in these periods. Since for the same period $[0, a^{f,i}]$, the two arrival processes are independent and for the second period, the first term is not affected, the independence is hence concluded. Consequently, the theorem follows from (47). ■

3) A further improved delay bound: In obtaining the improved delay bounds in Theorem 5, we made no assumption on the arrival process of the traversing flow or that of the crossing flow. If, however, these processes satisfy some assumptions, a further improved delay bound can be obtained.

Specifically, if $A^f(t)$ and $A^c(t)$ are independent and they have independent stationary increments, a further improved delay bound can be obtained.

Theorem 6. *Suppose that the traversing flow $A^f(t)$ and the crossing flow $A^c(t)$ are independent and they have independent stationary increments. Assume $M^f(1) \equiv E[e^{\theta A^f(1)}]$ and $M^c(1) \equiv E[e^{\theta A^c(1)}]$ exist for small $\theta > 0$ and $E[e^{\theta(A^f(1) + A^c(1) - C)}] \leq 1$. Then, for any packet $p^{f,i}$ of the traversing flow, its delay is bounded as*

$$P\{D^{f,i} > \tau\} \leq e^{-\theta(C - r^c)\tau}. \qquad (49)$$

for any $\theta \geq 0$ and any r^c such that $E[e^{\theta(A^c(1) - r^c)}] \leq 1$.

Proof: Our starting point is (35), which is reproduced here:

$$d^{f,i} \leq t^0 + \frac{\sum_{k=i_0}^{i} l^{f,k}}{C - r^c} + \frac{A^c(t^0, d^{f,i}) - r^c(d^{f,i} - t^0)}{C - r^c} \qquad (50)$$

with which, the following is easily verified

$$\begin{aligned} & (C - r^c) \cdot (d^{f,i} - a^{f,i}) \\ \leq\; & A^f(t^0, a^{f,i}) - (C - r^c) \cdot (a^{f,i} - t^0) \\ & + A^c(t^0, d^{f,i}) - r^c(d^{f,i} - t^0) \qquad (51) \\ =\; & A^f(t^0, a^{f,i}) + A^c(t^0, a^{f,i}) - C \cdot (a^{f,i} - t^0) \\ & + A^c(a^{f,i}, d^{f,i}) - r^c(d^{f,i} - a^{f,i}) \\ \leq\; & \sup_{0 \leq s \leq a^{f,i}} \{A^f(s, a^{f,i}) + A^c(s, a^{f,i}) - C \cdot (a^{f,i} - s)\} \\ & + A^c(a^{f,i}, d^{f,i}) - r^c(d^{f,i} - a^{f,i}) \qquad (52) \\ =\; & \sup_{0 \leq s \leq a^{f,i}} \big\{A^f(s, a^{f,i}) + A^c(s, a^{f,i}) - C \cdot (a^{f,i} - s) \\ & + A^c(a^{f,i}, d^{f,i}) - r^c(d^{f,i} - a^{f,i})\big\} \qquad (53) \end{aligned}$$

where in step (51) we have used the fact that $\sum_{k=i_0}^{i} l^{f,k} \leq A^f(t^0, a^{f,i})$.

It is worth highlighting that, the two terms in (52) are independent, since the second term is determined by a period that is non-overlapping with the period involved in the first term, and the process $A^c(t)$ has independent increments. Also due to this, in step (53), we have intentionally moved the second term inside $\sup\{\}$.

For ease of exposition, we let

$$Z = A^c(a^{f,i}, d^{f,i}) - r^c(d^{f,i} - a^{f,i})$$

for which, it is easily verified that, $E[e^{\theta Z}|d^{f,i}] = (E[e^{\theta(A^c(1)-r^c)}])^{d^{f,i}-a^{f,i}} \leq 1$ for $\forall d^{f,i}$ and hence $E[e^{\theta Z}] \leq 1$, under the given assumptions.

Then, for any $\theta \geq 0$, there holds,

$$\begin{aligned}
& P\{(C - r^c)D^{f,i} > x\} \\
= \;& P\{e^{\theta(C-r^c)(d^{f,i}-a^{f,i})} > e^{\theta x}\} \\
\leq \;& P\{e^{\sup_{0\leq s\leq a^{f,i}}\{\theta[A^f(s,a^{f,i})+A^c(s,a^{f,i})-C\cdot(a^{f,i}-s)]\}} \\
& \cdot e^{\theta[A^c(a^{f,i},d^{f,i})-r^c(d^{f,i}-a^{f,i})]} > e^{\theta x}\} \\
= \;& P\{\sup_{0\leq s\leq a^{f,i}} e^{\theta(A^f(s)+A^c(s)-C\cdot s)} \cdot e^{\theta Z} > e^{\theta x}\} && (54) \\
\leq \;& \frac{E[e^{\theta(A^f(1)+A^c(1)-C)}e^{\theta Z}]}{e^{\theta x}} && (55) \\
= \;& E[e^{\theta(A^f(1)+A^c(1)-C)}] \cdot E[e^{\theta Z}] \cdot e^{-\theta x} && (56) \\
\leq \;& e^{-\theta x} && (57)
\end{aligned}$$

where step (54) is due to that both $A^f(t)$ and $A^c(t)$ are stationary processes, step (56) is from the Doob's maximal inequalities for sub-(super-)martingales, and step (57) is from the assumptions of the theorem.

Specifically, define $X(s) = e^{\theta(A^f(s)+A^c(s)-C\cdot s)}e^{\theta Z}$, $s = 0, 1, 2, \ldots, a^{f,i}$. There holds, due to independent increments assumption,

$$\begin{aligned}
& E[X(s+1)|X(1), \ldots, X(s)] \\
= \;& E[e^{\theta(A^f(s,s+1)+A^c(s,s+1)-C)}]X(s) \\
= \;& E[e^{\theta(A^f(1)+A^c(1)-C)}]X(s) \\
\leq \;& X(s) && (58)
\end{aligned}$$

and hence $\{X(s)\}$ forms a supermartingale. Then (56) is obtained from the Doob's maximal inequality for supermartingales, which has also been used in the snetcal literature [6][18]. ■

VII. Examples

To demonstrate the obtained results, examples are presented in this section. The focus is on the obtained delay bounds. Without loss of generality and for ease of expression, we normalize the capacity and take $C = 1$.

A. Single Flow

For the single flow case, consider the arrival process $A^f(t)$ governed by a compound Poisson process. In this process, packets arrive according to a Poisson process with intensity λ. Packet lengths are independent and identically distributed, following a negative exponential distribution with mean $\frac{1}{\mu}$. Specifically:

$$A^f(t) = \sum_{n=1}^{N(t)} l^{f,i}$$

where $N(t)$ is a Poisson process with arrival intensity λ, which is independent of the packet lengths, and $l^{f,1}, l^{f,2}, \ldots$ are i.i.d. random variables with mean $\frac{1}{\mu}$.

For this compound Poisson process, it can be verified that it has a v.b.c. stochastic arrival curve [22] [19] $\alpha^f(t) = \frac{\lambda}{\mu-\theta}t$ with bounding function $\bar{F}^f(x) = e^{-\theta x}$ for $\forall \theta > 0$ and $r^f \equiv \frac{\lambda}{\mu-\theta} \leq 1$. Note that r^f here is a function of θ.

With Theorem 2, under the condition that $r^f \leq 1$, the tightest delay bound is obtained by taking $\theta = \mu - \lambda$, which is:

$$P\{D > \tau\} \leq e^{-(\mu-\lambda)\tau}. \tag{59}$$

It is worth highlighting that this single flow system may be considered[9] as an $M/M/1$ system with Poisson arrival rate λ and exponential service time distribution with parameter μ.

Appealingly, the delay bound (59) matches exactly with the delay[10] distribution found from $M/M/1$ analysis.

B. With Cross Traffic

For the case with cross traffic, we suppose that priority scheduling is adopted, with the crossing flow at the high priority level.

We assume the traversing flow and the crossing flow are independent of each other. For both, the arrival process is governed by a compound Poisson process. Similar to the single flow case, we consider that in each traffic arrival process, packets arrive according to a Poisson process with intensity λ^f for the traversing flow (respectively λ^c for the crossing flow). In addition, to ease later comparison, we assume all packets (of both flows) have the same i.i.d. length, following a negative exponential distribution with mean $1/\mu$.

This system is equivalent to an $M/M/1$/priority system, for which, the classic queueing theory has exact result for the delay expectation of the low priority traffic.

Note that, given the delay CCDF $P\{D \geq \tau\}$, the average delay is obtained as [24]

$$E[D] = \int_0^\infty P\{D \geq \tau\} d\tau.$$

[9] Note that in real computer networks the time is discrete. For this reason, we have also assumed discrete time at the beginning. Nevertheless, this paper does not specify the length of the time unit. Letting the unit time length → infinitely small, the system approaches time-continuous and all results in this paper still hold.

[10] It is the delay in the system which matches the definition of $D^{f,i}$, while not the delay in queue.

The above relationship between the delay expectation and the CCDF readily allows us to find upper bounds on delay expectation from the obtained delay bounds. Among the various delay bounds derived in the previous sections, (62) and (64) are the tightest and will be compared against the exact solution.

1) Delay expectation: For the $M/M/1/$priority system, the classic queueing theory gives the following result:

$$E[D] = \frac{\rho}{\mu(1-\rho^c)(1-\rho)} + \frac{1}{\mu} = \frac{1}{\mu(1-\rho)}[1 + \frac{\rho^c \rho}{1-\rho^c}] \quad (60)$$

where $E[D]$ denotes the delay expectation, $\rho^f \equiv \frac{\lambda^f}{\mu}$, $\rho^c \equiv \frac{\lambda^c}{\mu}$, and $\rho \equiv \rho^c + \rho^f$.

2) Bound on delay expectation, based on (44): Again, for the two compound Poisson arrival processes, the traversing flow has a v.b.c. stochastic arrival curve $\alpha^f(t) = \frac{\lambda^f}{\mu - \theta^f} t$ with bounding function $\bar{F}^f(x) = e^{-\theta^f x}$ for $\forall \theta^f > 0$; the traversing flow has a v.b.c. stochastic arrival curve $\alpha^c(t) = \frac{\lambda^c}{\mu - \theta^c} t$ with bounding function $\bar{F}(x) = e^{-\theta^c x}$ for $\forall \theta^c > 0$.

For ease of expression, letting $\theta^f = \theta^c \equiv \theta$, which may give a sub-tight bound, we obtain from Theorem 5

$$P\{D > \tau\} \leq (1 + \theta \cdot (1 - r^c) y) e^{-\theta \cdot (1 - r^c)\tau} \quad (61)$$

where $r^c = \frac{\lambda^c}{\mu - \theta}$ and $r^f = \frac{\lambda^f}{\mu - \theta}$, for any $\theta > 0$, satisfying

$$r^f + r^c \leq 1$$

which further gives, by letting $\theta = \mu - \lambda^f - \lambda^c$

$$P\{D > y\} \leq [1 + \frac{\lambda^f (1-\rho)}{\rho} y] e^{-\frac{\lambda^f (1-\rho)}{\rho} y}$$

and consequently, a bound on delay expectation is as:

$$E[D] \leq \frac{2}{\mu(1-\rho)}[1 + \frac{\rho^c}{\rho^f}]. \quad (62)$$

3) Bound on delay expectation, based on (49): For the considered system, letting $\theta = \mu - \lambda^f - \lambda_c$ and $r^c = \frac{\lambda^c}{\lambda^f + \lambda^c}$, the following can be verified: (1) $E[e^{\theta(A^c(1) - r^c)}] = e^{\theta(\frac{\lambda^c}{\mu - \theta} - r^c)} = 1$ and (2) $E[e^{\theta(A^f(1) + A^c(1) - C)}] = e^{\theta(\frac{\lambda^f}{\mu - \theta} + \frac{\lambda^c}{\mu - \theta} - 1)} = 1$. Then, from Theorem 6, the delay bound (49) becomes

$$P\{D > \tau\} \leq e^{-(\mu - \lambda^f - \lambda_c)(1 - r^c)\tau} \quad (63)$$

with which, the following bound on delay expectation is obtained:

$$E[D] \leq \frac{1}{\mu(1-r^c)(1-\rho)} = \frac{1}{\mu(1-\rho)}[1 + \frac{\rho^c}{\rho^f}] \quad (64)$$

which is clearly better than (62).

To give an overview of the bound (64), Fig. 1 is presented, where x-axis is the total load, y-axis is the share of cross traffic in the total load. In the figure, the bound is compared against the exact result (60), under different total loads ($\rho \in [0, 0.9]$), and different shares ($\frac{\rho^c}{\rho} \in [0, 0.9]$). The comparison shows that the bound (64) is reasonably good.

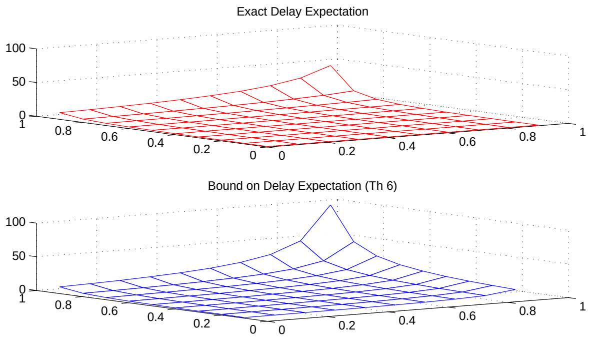

Fig. 1. Comparison of bound (64) ($\mu = 1$)

VIII. Discussion and Related Work

In deterministic network calculus, the delay bound derived from the Guaranteed Rate server model is better than that directly from the deterministic counterpart of (8). To overcome this difference, an interesting property has been proved, which says, in deterministic delay bound analysis, *the last packetizer can be ignored* [5][27]. For the considered single node case, this property implies that, for the concern of deterministic delay bound analysis, the constant capacity node could be treated as if it had a deterministic service curve $C \cdot t$ and hence Proposition 3 could be used directly. Results in this paper further imply that this property can also be extended to the stochastic network calculus context. Particularly, it is easily verified that, for the single flow case, delay bound (23) in Theorem 2 is better than delay bound (22) in Corollary 1 by ignoring the packetization effect $\bar{F}_l$. In addition, for the case with cross traffic, Corollary 2 and Corollary 3 will lead to Theorem 6 by ignoring the packetization effect.

In the general sense of taking packetization effect into stochastic service curve and delay bound analysis, the work [2] is most related. However, the obtained results in [2] are mostly functions of $\int_x^\infty \bar{F}^l(y)dy$, while in our results, they are related directly to $\bar{F}^l$. In addition, how to make use of independence information to improve the obtained results is not investigated in [2]. Moreover, [2] focuses on a specific type of traffic, while our investigation is more systematic (for the single node case), applicable to any type of traffic that has v.b.c stochastic arrival curve, which covers a wide range of traffic types [22].

For the examples, delay bound analysis of $M/M/1$ using snetcal can be found in [6][18]. However, the technique used in this paper has fundamental difference from the techniques used in [6][18]. Particularly in [6][18], the analysis directly works on the arrival process and the service process, without mapping the arrival process to the stochastic arrival curve characterization, nor proving the stochastic service curve characterization of the system taking into consideration the packetization effect. For delay bound analysis of $M/M/1/$priority using snetcal, the same delay expectation bound as (64) may be found in [7]. However, beside the fundamental difference in the used analytical technique, the bound in [7] is derived under some additional conditions/assumptions, e.g., preemptive priority and ignoring the packetizer. Nevertheless, it is exciting to see the same bound derived when the packetization effect is taken into account.

IX. Conclusion

In this paper, we considered a packet-switched network node with constant capacity (in bps) and systematically derived stochastic service curves and delay bounds for the system. Specifically, we proved that the node provides a stochastic service curve with a bounding function equal to the CCDF of packet length distribution. In addition, we derived delay bounds, which imply that *the last packetizer can be ignored* property may be extended to SNC. Furthermore, we presented relations that allow to exploit independence information in the analysis. For the single flow case, a by-product is a new delay bound that matches with the exact result for $M/M/1$.

Recall that, while the considered system is perhaps the simplest computer network system, before this work, in the context of stochastic network calculus, little was known about how to make use of the independence information in the analysis, particularly when the packetization effect is considered. This paper makes one step forward. We believe the analysis may be extended to the network case, where how to make use of flow independence information to improve results (without ignoring the packetization effect) still remains largely mysterious.

References

[1] S. Blake and et al. An architecture for Differentiated Services. *IETF RFC 2475*, Dec. 1998.

[2] A. Burchard, J. Liebeherr, and F. Ciucu. On superlinear scaling of network delays. *IEEE/ACM Trans. Netw.*, 19(4):1043–1056, 2011.

[3] A. Burchard, J. Liebeherr, and S. D. Patek. A min-plus calculus for end-to-end statistical service guarantees. *IEEE Trans. Information Theory*, 52:4105–4114, 2006.

[4] C. S. Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Trans. Auto. Control*, 39(5):913–931, May 1994.

[5] C.-S. Chang. *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000.

[6] F. Ciucu. Network calculus delay bounds in queueing networks with exact solutions. In *Proc. ITC 2007*, 2007.

[7] F. Ciucu. *Scaling Properties in the Stochastic Network Calculus*. PhD thesis, University of Virginia, August 2007.

[8] F. Ciucu, A. Burchard, and J. Liebeherr. A network service curve approach for the stochastic analysis of networks. *IEEE Trans. Information Theory*, 52(6):2300–2312, June 2006.

[9] F. Ciucu and J. Schmitt. Perspectives on network calculus: no free lunch, but still good value. In *SIGCOMM*, pages 311–322, 2012.

[10] R. L. Cruz. A calculus for network delay, part I and part II. *IEEE Trans. Information Theory*, 37(1):114–141, Jan. 1991.

[11] R. L. Cruz. Quality of service management in integrated services networks. In *Proc. 1st Semi-Annual Research Review, CWC, UCSD*, June 1996.

[12] M. Fidler. An end-to-end probabilistic network calculus with moment generating functions. In *Proceedings of IWQoS*, 2006.

[13] M. Fidler. A survey of deterministic and stochastic service curve models in the network calculus. *IEEE Comm. Sur. & Tut.*, 12(1), 2010.

[14] Y. Jiang. Delay bounds for a network of Guaranteed Rate servers with FIFO aggregation. *Computer Networks*, 40(6):683–694, Dec. 2002.

[15] Y. Jiang. Relationship between guaranteed rate server and latency rate server. *Computer Networks*, 43(3):307–315, 2003.

[16] Y. Jiang. A basic stochastic network calculus. In *Proc. ACM SIGCOMM 2006*, pages 123–134, 2006.

[17] Y. Jiang. Per-domain packet scale rate guarantee for expedited forwarding. *IEEE/ACM Trans. Networking*, 14:630–643, 2006.

[18] Y. Jiang. Network calculus and queueing theory: Two sides of one coin. In *Proc. Valuetools*, 2009. Updated version on 16-March-2010.

[19] Y. Jiang. A note on applying stochastic network calculus, May 2010.

[20] Y. Jiang. Stochastic network calculus for performance analysis of internet networks: An overview and outlook. In *Proc. ICNC*, 2012.

[21] Y. Jiang and Y. Liu. *Stochastic Network Calculus*. Springer-Verlag, 2008.

[22] Y. Jiang, Q. Yin, Y. Liu, and S. Jiang. Fundamental calculus on generalized stochastically bounded bursty traffic for communication networks. *Computer Networks*, 53(12):2011–2021, 2009.

[23] F. Kelly. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications, Royal Statistical Society Lecture Notes Series, 4. Oxford University Press*, 1996.

[24] L. Kleinrock. *Queueing Systems, Volume 1: Theory*. John Wiley & Sons, 1975.

[25] A. Kumar, D. Manjunath, and J. Kuri. *Communication Networking: An Analytical Approach*. Morgan Kaufmann, 2004.

[26] J. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *ACM SIGMETRICS'92*, 1992.

[27] J.-Y. Le Boudec and P. Thiran. *Network Calculus: A Theory of Deterministic Queueing Systems for the Internet*. Springer-Verlag, 2001.

[28] K. Lee. Performance bounds in communication networks with variable-rate links. In *Proc. ACM SIGCOMM'95*, 1995.

[29] C. Li, A. Burchard, and J. Liebeherr. A network calculus with effective bandwidth. *IEEE/ACM Trans. Networking*, 15(6):1442–1453, December 2007.

[30] J. Liebeherr, A. Burchard, and F. Ciucu. Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Transactions on Information Theory*, 58(2):1010–1024, 2012.

[31] Y. Liu, C.-K. Tham, and Y. Jiang. A calculus for stochastic QoS analysis. *Performance Evaluation*, 64:547–572, 2007.

[32] S. Mao and S. S. Panwar. A survey of envelope processes and their applications in quality of service provisioning. *IEEE Comm. Sur. & Tut.*, 8(1-4):2–20, 2006.

[33] O. Yaron and M. Sidi. Performance and stability of communication network via robust exponential bounds. *IEEE/ACM Trans. Networking*, 1(3):372–385, June 1993.

[34] 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.

Appendix: Proof of Lemma 1

Consider any time t and any sample path of the system. Since $i(t) = \min\{k : d^{f,k} \geq t\}$, meaning $p^{f,i(t)}$ is the most recent packet of flow f in $A^*(t)$ which departs from the node after t, there holds:

$$A^*(t) \geq \sum_{k=1}^{i(t)-1} l^{f,k} \tag{65}$$

where the equality holds when $d^{f,i(t)} > t$; otherwise when $d^{f,i(t)} = t$, $A^*(t) > \sum_{k=1}^{i(t)-1} l^{f,k}$ since $A^*(t) = \sum_{k=1}^{i(t)} l^{f,k}$.

Define

$$i' = \max\{k : a^{f,k} \leq t\}$$

which means that $p^{f,i'}$ is the last arrival packet in $A(t)$ which arrives at or before time t. In other words, we have

$$a^{f,i'+1} > t.$$

In addition, comparing i' with $i(t)$, we must have

$$i' \geq i(t) - 1$$

because $i(t)-1$ is the last departure packet before or on time t, which has to have arrived before or at time t.

Let us split the time period $[0,t]$ into $i'+1$ intervals, which are $[0,a^{f,1})$, …, $[a^{f,j-1},a^{f,j})$, …, $[a^{f,i'-1},a^{f,i'})$ and $[a^{f,i'},t]$. Then, consider

$$\begin{aligned}
& A\otimes\beta_h(t)-A^*(t) \\
= \;& \inf_{0\le s\le t}\{A(s)+R\cdot(t-s)-A^*(t)\} \\
= \;& R\cdot\inf_{0\le s\le t}\{(t-s)+\frac{A(s)-A^*(t)}{R}\} \\
= \;& R\cdot\min\Bigg[\inf_{0\le s<a^{f,1}}\{(t-s)+\frac{A(s)-A^*(t)}{R}\},\ldots, \\
& \inf_{a^{f,j-1}\le s<a^{f,j}}\{(t-s)+\frac{A(s)-A^*(t)}{R}\},\ldots, \\
& \inf_{a^{f,i'}\le s\le t}\{(t-s)+\frac{A(s)-A^*(t)}{R}\}\Bigg]
\end{aligned}\tag{66}$$

For the first interval, we have $A(s)=0$ for $\forall s: 0\le s<a^{f,1}$. Hence

$$\begin{aligned}
& \inf_{0\le s<a^{f,1}}\{(t-s)+\frac{A(s)-A^*(t)}{R}\} \\
= \;& t-\frac{A^*(t)}{R}+\inf_{0\le s<a^{f,1}}\{-s+\frac{A(s)}{R}\} \\
= \;& t-\frac{A^*(t)}{R}-a^{f,1} \\
\le \;& d^{f,i(t)}-\left[a^{f,1}+\frac{A^*(t)}{R}\right]
\end{aligned}\tag{67}$$

Similarly, for the next $i'-1$ intervals, we have $A(s)=\sum_{k=1}^{j-1}l^{f,k}$, for $\forall s: a^{f,j-1}\le s<a^{f,j}$, $(j=2,\ldots,i')$. Hence,

$$\begin{aligned}
& \inf_{a^{f,j-1}\le s<a^{f,j}}\{(t-s)+\frac{A(s)-A^*(t)}{R}\} \\
= \;& t-\frac{A^*(t)}{R}+\inf_{a^{f,j-1}\le s<a^{f,j}}\{-s+\frac{A(s)}{R}\} \\
= \;& t-\frac{A^*(t)}{R}-a^{f,j}+\frac{\sum_{k=1}^{j-1}l^{f,k}}{R} \qquad (68) \\
\le \;& d^{f,i(t)}-\left[a^{f,j}+\frac{A^*(t)-\sum_{k=1}^{j-1}l^{f,k}}{R}\right] \qquad (69)
\end{aligned}$$

Note that for $\forall s: a^{f,i'}\le s<a^{f,i'+1}$, we have $A(s)=\sum_{k=1}^{i'}l^{f,k}$. In addition, in the above discussion, it is known $t<a^{f,i'+1}$. Hence, for the last interval, we have $A(s)=\sum_{k=1}^{i'}l^{f,k}$ for $\forall s: a^{f,i'}\le s\le t$. Consequently,

$$\begin{aligned}
& \inf_{a^{f,i'}\le s\le t}\{(t-s)+\frac{A(s)-A^*(t)}{R}\} \\
= \;& t-\frac{A^*(t)}{R}+\inf_{a^{f,i'}\le s\le t}\{-s+\frac{A(s)}{R}\} \\
= \;& t-\frac{A^*(t)}{R}-t+\frac{\sum_{k=1}^{i'}l^{f,k}}{R} \\
= \;& -\frac{A^*(t)}{R}+\frac{\sum_{k=1}^{i'}l^{f,k}}{R} \\
\le \;& d^{f,i(t)}-\left[a^{f,i(t)}+\frac{A^*(t)-\sum_{k=1}^{i'}l^{f,k}}{R}\right]
\end{aligned}\tag{70}$$

Considering all these $i'+1$ intervals, we get

$$\begin{aligned}
& A\otimes\beta_h(t)-A^*(t) \\
\le \;& R\cdot d^{f,i(t)}- \\
& R\cdot\max[a^{f,1}+\frac{A^*(t)}{R},\ldots,a^{f,j}+\frac{A^*(t)-\sum_{k=1}^{j-1}l^{f,k}}{R},\ldots, \\
& a^{f,i'}+\frac{A^*(t)-\sum_{k=1}^{i'}l^{f,k}}{R},a^{f,i(t)}+\frac{\sum_{k=i'+1}^{i(t)}l^{f,k}}{R}] \qquad (71) \\
\le \;& l^{f,i(t)}+R\cdot d^{f,i(t)}- \\
& R\cdot\max[a^{f,1}+\frac{\sum_{k=1}^{i(t)}l^{f,k}}{R},\ldots,a^{f,j}+\frac{\sum_{k=j}^{i(t)}l^{f,k}}{R},\ldots, \\
& a^{f,i'}+\frac{\sum_{k=i'}^{i(t)}l^{f,k}}{R},a^{f,i(t)}+\frac{\sum_{k=i'+1}^{i(t)}l^{f,k}}{R}] \qquad (72) \\
\le \;& l^{f,i(t)}+R\cdot d^{f,i(t)}- \\
& R\cdot\max\left[a^{f,1}+\frac{\sum_{k=1}^{i(t)}l^{f,k}}{R},\ldots,a^{f,i(t)}+\frac{\sum_{k=i(t)}^{i(t)}l^{f,k}}{R}\right] \qquad (73) \\
= \;& R\cdot[d^{f,i(t)}-V^{f,i(t)}(R)]+l^{f,i(t)}. \qquad (74)
\end{aligned}$$

Here step (73) is due to $i(t)\le i'+1$ and taking maximum on the first $i(t)$ elements of the third term in (72) results in a smaller or equal value.