Network Calculus Delay Bounds in Multi-Server Queueing Networks with Stochastic Arrivals and Stochastic Services

Zhidu Li, Yuehong Gao, Bala Alhaji Salihu, Pengxiang Li, Lin Sang, Dacheng Yang

Wireless Theories and Technologies Lab
Beijing University of Posts and Telecommunications
Beijing, P.R. China, 100876
Email: prclzd@126.com

***Abstract*—Studies of multi-server networks are usually conducted on the assumptions of independent servers and specific arrivals, which may not capture the characteristics of realistic networks. In this paper, we propose a novel stochastic network calculus approach to perform delay analysis for a general multi-server network. The stochastic network service curves are derived in two ways with different assumptions. After that, we use these service curves to derive queueing delay bounds including delay bound distribution and mean delay bound for a specific scenario of which the classical M/M/N model is a subset. Compared with previous studies, it is worth highlighting that our analysis contains the cases where the arrival and service processes as well as the servers can either be independent or correlated. In addition, the accuracy of the analytical delay bounds are verified by comparing with the queueing theory results in an M/M/N model.**

Keywords*—*Multi-server network; stochastic service curve; delay bound; stochastic network calculus

I. Introduction

Multi-server network refers to a network with multiple parallel servers or channels which can transmit data simultaneously. Examples include LTE-OFDM (Long Term Evolution-orthogonal frequency division multiplexing) networks in which the available bandwidth is divided into hundreds of orthogonal channels (subcarriers) to be shared by different users [1]. QoS (quality of service) for multi-server networks is an active area of research in both academia and industry. Particularly, delay is one of the key performance metrics being considered.

Traditional queueing theory is a classical approach to deal with delay problems in multi-server networks [2]. Retrospecting previous researches, delay analysis based on queueing theory have been performed in several multi-server networks. For instance, Lu used an M/M/N model to derive mean delay based on his traffic scheduling algorithm in hybrid fiber-coaxial networks [3]. Delay analysis in MAC (medium access control) layer with M/M/N queueing model can be found in [4, 5]. However, queueing theory has two drawbacks. One is the statistically independent assumption between the arrival and service processes and amongst the servers. The other drawback is that the results of queueing theory are usually presented as mean values rather than probabilistic distributions especially when network structure is complicate. Hence, queueing theory based analysis is usually intractable for realistic multi-server networks [6].

In this paper, we propose a stochastic network calculus approach to analyze the queueing delay for a one queue multi-server network of which the queueing models of [3–5] are all subsets. Specifically, the paper considers networks with arbitrary number of parallel servers, stochastic arrival traffic and time varying service in MAC layer. In particular, the arrival and service processes can either be independent or correlated, and so are the servers. After deriving the stochastic service curve for a general network model, we perform delay analysis in a specific scenario which will reduce to an M/M/N model if arrival and service processes as well as servers are statistically independent. The delay distributions and mean delay bounds are derived for different cases. Moreover, for verifying the accuracy of our analysis, we compare our analytical results with the comparable queueing theory results through numerical simulations.

As an analytical tool employed in this paper, stochastic network calculus [7, 8] has been proved to perform well in performance analysis for single server networks and tandem server networks (e.g. [9–14] etc.). Its core idea is to use arrival curve and service curve to yield performance bounds (e.g. delay and backlog bounds). While dealing with problems, the results of stochastic network calculus are usually presented as probabilistic bounds, which in turn simplifies the complex nature of non-linear problems albeit at the expense of some degree of accuracy. Further, it has been shown that the stochastic network calculus results are in accordance with the queueing theory results in M/M/1 and M/D/1 queueing models [13, 14].

In the literature, the limits of queueing theory in delay analysis have been highlighted in the second paragraph above. To the best of our knowledge, researches of multi-server networks based on stochastic network calculus are few. Usually, these works were carried out on the assumption of statistically independent servers with constant service rate, such as the works in [15, 16]. Additionally, analysis based on other theories can be found in [17–19]. In such works, specific arrivals (e.g. Bernoulli arrivals) and independent servers with

This work is supported by the National Science Foundation of China under Grant No.61300185.

978-1-4799-5952-5/15/$31.00 ©2015 IEEE

constant service rate were often assumed.

Therefore, compared with the previous works, the main contributions of this paper are as follows:

- We extend the theory of stochastic network calculus to a multi-server framework which is applied to arbitrary number of servers, stochastic arrival process and stochastic service process. Significantly, we synthetically consider the interdependence between the arrival and service processes and amongst the servers. Thus, our work is more close to the realistic scenario.
- We derive stochastic service curves of the whole network and delay bounds for different cases. Particularly, the delay bound distributions are infeasible in queueing theory when the arrival and service processes or the serves are not independent.

The rest of the paper is structured as follows. Section II describes a general multi-server network model and the analytical tool. In Section III, we first derive the stochastic network service curves for different cases. After that, we conduct delay bound analysis in a specific scenario. In Section IV, numerical analysis is presented. Finally, Section V concludes the paper.

II. System Model and Network Calculus Basics

A. System Model

We consider a multi-server network model containing N servers, infinite waiting buffer and a scheduler as depicted in Fig. 1. A server is meant to serve only one packet at a time and a packet can be served by only one server. The arrival packets are served in the principle of first-in-first-out (FIFO). The scheduler will allocate a server to serve the head packet of the buffer based on a given mechanism (e.g. load balance, priority servers etc.) provided that there are any idle servers. If all the servers are busy, the packets would have to wait in the buffer.

In this paper, we can use different stochastic processes to characterize the services of different servers and the arrival traffic respectively. We use $S(s,t)$ to denote the cumulative service guaranteed by the whole network within time interval $(s,t]$, and $S(t) = S(0,t)$, such that we have $S(s,t) = S(t) - S(s)$. The server indexed i $(1 \leq i \leq N)$ is denoted by S_i. The cumulative service guaranteed by S_i up to time t is denoted by $S_i(t)$. Similarly, we denote the cumulative arrivals and departures of the network by $A(t)$ and $A^*(t)$ respectively. The arrival process and departure processes for server S_i are denoted by $A_i(t)$ and $A_i^*(t)$ respectively. In addition, the buffer and the cumulative processes for the network are all set to 0 at time 0, thus, $A(0) = A^*(0) = S(0) = 0$.

B. Stochastic Network Calculus Basics

There are two basic concepts in stochastic network calculus: the stochastic arrival curve (SAC) and the stochastic service curve (SSC), which are used to describe the arrival process of input traffic and the service process of server respectively. For ease of understanding, we introduce basic definitions and theorem which are useful to the subsequent analysis [8].

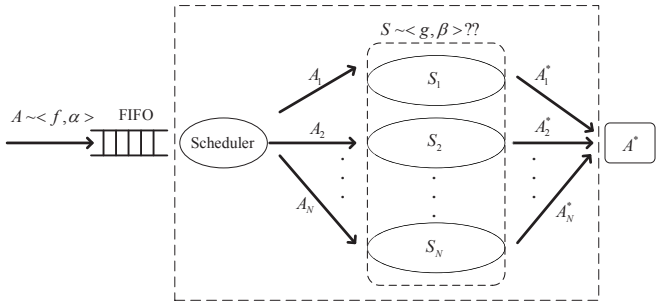

Fig. 1. Multi-server network model

Definition 1. ***(Stochastic Arrival Curve)*** *A flow A is said to have a stochastic arrival curve $\alpha \in \Psi$[1] with the bounding function $f \in \bar{\Psi}$[1], denoted by $A \sim< f, \alpha >$, if for all $t \geq s \geq 0$ and all $x \geq 0$, there holds*

$$Pr\{\sup_{0 \leq s \leq t} \{A(s,t) - \alpha(s,t)\} > x\} \leq f(x).$$

Definition 2. ***(Stochastic Service Curve)*** *A system S is said to provide a stochastic service curve $\beta \in \Psi$ with the bounding function $g \in \bar{\Psi}$, denoted by $S \sim< g, \beta >$, if for all $t \geq 0$ and all $x \geq 0$, there holds*

$$Pr\{A \otimes \beta(t) - A^*(t) > x\} \leq g(x),$$

where operator $\otimes$ represents the min-plus convolution, and

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

Theorem 1. ***(Delay Bound Distribution)*** *Consider a system S with input A. Suppose A has a SAC as $A \sim< f, \alpha >$; and server S provides service with a SSC as $S \sim< g, \beta >$. Then, for all $t \geq 0$ and $x \geq 0$, no matter whether the arrival process is independent of the service process or not, the delay $W(t)$ is bounded by:*

$$Pr\{W(t) > h(\alpha + x, \beta)\} \leq [f \otimes g(x)]_1 \quad , \tag{1}$$

where function $[\cdot]_1$ represents min$\{\cdot, 1\}$, and $h(\alpha + x, \beta)$ holds as follows

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

Particularly, if the arrival process is independent of the service process, there holds

$$\begin{aligned} &Pr\{W(t) > h(\alpha + x, \beta)\} \\ \leq &1 - \int_{-\infty}^{+\infty} (1 - f(x - y)) d(1 - g(y)) \end{aligned} \quad . \tag{2}$$

For using Theorem 1 to ascertain the delay bound distribution of the multi-server network, the key challenge is to derive an equivalent SSC of the whole network. In this paper, we call this equivalent SSC the stochastic network service curve and use $S \sim< g, \beta >$ to denote it. Also, the delay bound derived in this paper is an upper bound of the queueing delay.

[1] In this paper, we use Ψ and $\bar{\Psi}$ to denote the set of non-negative wide sense increasing and decreasing function respectively.

III. Performance Analysis

In this section, the stochastic network service curve will be derived. Thereafter, we will perform delay bound analysis for a specific scenario of which the classical M/M/N system is a subset.

A. Stochastic Network Service Curve

The stochastic network service curve will be derived in two ways and formulated as Theorem 2 and Theorem 3. In Theorem 2, the service process of each server is not required to have finite moment generating function (i.e., the SSC of each server could be derived in other ways other than the limited terms of [20]), however, the information of the SAC and the SSC for each server is needed. In Theorem 3, we derive the stochastic network service curve without any information of the arrivals on condition that the service process of each server has finite moment generating function and stationary increments.

Before deriving the stochastic network service curve, the following lemma is needed (proof can be found in [8]).

Lemma 1. *For the sum of a collection of random variables $Z = \sum_{i=1}^{N} X_i$, no matter whether they are independent or not, there holds for the CCDF (complementary cumulative distribution function) of Z*

$$\bar{F}_Z(z) \leq [\bar{F}_{X_1} \otimes \bar{F}_{X_2} \otimes \cdots \otimes \bar{F}_{X_N}(z)]_1$$

Theorem 2. *Consider a multi-server network S with N servers $\{S_i | 1 \leq i \leq N\}$. For $\forall i$, the SSC of S_i is given by $S_i \sim< g_i, \beta_i >$, Suppose a stochastic arrival traffic $A \sim< f, \alpha >$ is divided into N arrival processes as $\{A_i \sim< f_i, \alpha_i > | 1 \leq i \leq N\}$ to be served by S_i respectively based on a given scheduling principle. The whole network then guarantees a stochastic network service curve $S \sim< g, \beta >$ with*

$$\beta(t) = \sum_{i=1}^{N} \beta_i(t)$$

$$g(x) = [f_1 \cdots \otimes f_N \otimes g_1 \otimes \cdots \otimes g_N(x - \sum_{i=1}^{N} \alpha_i \oslash \beta_i(0))]_1$$

where $\alpha_i \oslash \beta_i(0) = \sup_{s \geq 0}\{\alpha(s) - \beta(s)\}$.

Proof: In light of the description of Section II-A, the network arrivals can be regarded as the aggregation of the arrivals for each server. Hence, we have $A(t) = \sum_{i=1}^{N} A_i(t)$. Similar relationship between the departure processes can be expressed as $A^*(t) = \sum_{i=1}^{N} A_i^*(t)$.

We first find out the upper bound of the expression $A \otimes \beta(t) - \sum_{i=1}^{N} A_i \otimes \beta_i(t)$. We have

$$\begin{aligned}
&A \otimes \beta(t) - \sum_{i=1}^{N} A_i \otimes \beta_i(t) \\
&= \inf_{0 \leq s \leq t}\{A(s) + \beta(s,t)\} - \sum_{i=1}^{N} \inf_{0 \leq s \leq t}\{A_i(s_i) + \beta_i(s_i,t)\} \\
&\leq \sum_{i=1}^{N}(A_i(t) + \beta_i(t,t)) - \sum_{i=1}^{N} \inf_{0 \leq s \leq t}\{A_i(s_i) + \beta_i(s_i,t)\} \\
&= \sum_{i=1}^{N} \sup_{0 \leq s_i \leq t}\{A_i(s_i,t) - \alpha_i(s_i,t) + \alpha_i(s_i,t) - \beta_i(s_i,t)\} \\
&\leq \sum_{i=1}^{N}(\sup_{0 \leq s_i \leq t}\{A_i(s_i,t) - \alpha_i(s_i,t)\} \\
&\qquad + \sup_{0 \leq s_i \leq t}\{\alpha_i(s_i,t) - \beta_i(s_i,t)\}) \\
&\leq \sum_{i=1}^{N} \sup_{0 \leq s_i \leq t}\{A_i(s_i,t) - \alpha_i(s_i,t)\} + \sum_{i=1}^{N} \alpha_i \oslash \beta_i(0)
\end{aligned}$$

Applying Definition 2, we have

$$\begin{aligned}
&Pr\{A \otimes \beta(t) - A^*(t) > x\} \\
&\leq Pr\{\sum_{i=1}^{N}(A_i \otimes \beta_i(t) + \sup_{0 \leq s_i \leq t}\{A_i(s_i,t) - \alpha_i(s_i,t)\} \\
&\qquad + \alpha_i \oslash \beta_i(0) - A_i^*(t)) > x\} \\
&\overset{(a)}{=} Pr\{\sum_{i=1}^{N}(\sup_{0 \leq s_i \leq t}\{A_i(s_i,t) - \alpha_i(s_i,t)\} \\
&\qquad + A_i \otimes \beta_i(t) - A^*(t)) > x - \sum_{i=1}^{N} \alpha_i \oslash \beta_i(0)\} \\
&\overset{(b)}{\leq} [f_1 \otimes \cdots \otimes f_N \otimes g_1 \otimes g_2 \otimes \cdots g_N(x - \sum_{i=1}^{N} \alpha_i \oslash \beta_i(0))]_1
\end{aligned}$$

where the steps from (a) to (b) are based on Definition 1 and Lemma 1. Therefore, Theorem 2 is proved. ■

Significantly, the stochastic network service curve in Theorem 2 is applied to arbitrary service processes as long as the SSC of each server can be obtained. However, it is correlated with the arrivals. If we augment an assumption that for all $1 \leq i \leq N$, $S_i(t)$ has finite moment generating function and stationary increments, the arrival and service processes will be decoupled.

The moment generating function of a stochastic process $X(t)$ is defined as $\mathrm{E}[e^{\theta X(t)}]$ [7]. Moreover, if $\log \mathrm{E}[e^{\theta X(t)}]/t \leq \infty$, we say $X(t)$ has finite moment generating function.

Theorem 3. *Consider a multi-server network S with N servers $\{S_i | 1 \leq i \leq N\}$. If $\forall i$, $S_i(t)$ has finite moment generating function and stationary increments, then no matter whether the servers are independent or not, for all $t \geq 0$ and $x \geq 0$, the whole network guarantees a stochastic network service curve $S \sim< g, \beta >$ with*

$$\begin{aligned}
\beta(t) &= \sum_{i=1}^{N} \beta_i(t) = \sum_{i=1}^{N}(-\frac{1}{\theta}\log E[e^{-\theta S_i(t)}]) \\
g(x) &= [N e^{-\frac{\theta x}{N}}]_1
\end{aligned} \quad . \tag{3}$$

Here, $\beta_i(t) = -\frac{1}{\theta}\log E[e^{-\theta S_i(t)}]$, denotes the SSC of S_i, and θ is a nonnegative free parameter. In particular, if all the servers are independent, a tighter bounding function holds as follows

$$g(x) = e^{-\theta x}. \tag{4}$$

Proof: Let us consider any time $t \geq 0$. There are two cases.

Case 1: t is not in any backlogged period. In this case, there is no backlog in the network at time t, which means all the traffic arriving up to time t has left the network. Hence, we have $A(t) = A^*(t)$, and

$$A \otimes \beta(t) - A^*(t) = \beta(t,t) - S(t,t).$$

Case 2: t is within a backlogged period. Suppose t_0 is the start time of the backlogged period and $t_0 \leq t$. Thus, we have $A(t_0) = A^*(t_0)$, and

$$\begin{aligned} &A \otimes \beta(t) - A^*(t) \leq A(t_0) + \beta(t_0,t) - A^*(t) \\ =&\beta(t_0,t) - A^*(t_0,t) = \beta(t_0,t) - S(t_0,t) \\ =&\sum_{i=1}^{N}(\beta_i(t_0,t) - S_i(t_0,t)) \end{aligned}.$$

Thus, for all $x > 0$, we have

$$\begin{aligned} &Pr\{A \otimes \beta(t) - A^*(t) > x\} \\ \leq &Pr\{\sum_{i=1}^{N}(\beta_i(t_0,t) - S_i(t_0,t)) > x\} \end{aligned}.$$

Then for the service processes $\{S_i(t)|1 \leq i \leq N\}$ with finite moment generating functions and stationary increments, no matter whether the servers are independent or not, there holds

$$\begin{aligned} &Pr\{A \otimes \beta(t) - A^*(t) > x\} \\ \leq &Pr\{\sum_{i=1}^{N}(\beta_i(t_0,t) - S_i(t_0,t))) > x\} \\ \overset{(a)}{\leq} &\inf_{x_1+\ldots+x_N=x}\{\sum_{i=1}^{N}(Pr\{\beta_i(t_0,t) - S_i(t_0,t)) > x_i\})\} \\ \overset{(b)}{\leq} &\inf_{x_1+\ldots+x_N=x}\{\sum_{i=1}^{N}(e^{-\theta x_i}\mathrm{E}[e^{(\theta(\beta_i(t_0,t)-S_i(t_0,t))}])\} \\ = &\inf_{x_1+\ldots+x_N=x}\{\sum_{i=1}^{N}(e^{-\theta x_i}\mathrm{E}[e^{(\theta(\beta_i(\tau)-S_i(\tau))}])\} \\ \overset{(c)}{=} &\inf_{x_1+\ldots+x_N=x}\{\sum_{i=1}^{N}e^{-\theta x_i}\} \\ =&[Ne^{-\frac{\theta x}{N}}]_1 \end{aligned}.$$

Here, in (a) we apply Lemma 1 and in (b) we apply the Chernoff bound. In (c) we let $\beta_i(t) = -\frac{1}{\theta}\log \mathrm{E}[e^{-\theta S_i(t)}]$.

On the other hand, if all the servers are statistically independent, we have;

$$\begin{aligned} &Pr\{A \otimes \beta(t) - A^*(t) > x\} \\ \leq &Pr\{\sum_{i=1}^{N}(\beta_i(t_0,t) - S_i(t_0,t)) > x\} \\ \overset{(a)}{\leq} &e^{-\theta x}\mathrm{E}[e^{\theta\sum_{i=1}^{N}(\beta_i(t_0,t)-S_i(t_0,t))}] \\ \overset{(b)}{\leq} &e^{-\theta x}\prod_{i=1}^{N}\mathrm{E}[e^{\theta(\beta_i(t_0,t)-S_i(t_0,t))}] \\ =&e^{-\theta x}\prod_{i=1}^{N}\mathrm{E}[e^{\theta(\beta_i(\tau)-S_i(\tau))}] \\ \overset{(c)}{=}&e^{-\theta x} \end{aligned}.$$

Here, in (a) we applied the Chernoff bound. Step (b) holds since the servers are independent. In (c), we let $\beta_i(t) = -\frac{1}{\theta}\log \mathrm{E}[e^{-\theta S_i(t)}]$. Hence, Theorem 3 is proved. ■

Actually, the assumptions of finite moment generating function and stationary increments in Theorem 3 are logical, because many types of stochastic processes have these characteristics, such as Poisson process, exponential on-off process, Markov modulated process etc. [20].

B. Delay Bound Analysis for Specific Scenario

In this subsection, we consider a specific multi-server network scenario as follows. The number of arrival packets follows a Poisson distribution with mean rate λ (packets/s). The packet size is fixed to L (bits). The service rate of each server is time varying and the service time is exponentially distributed with mean time $1/\mu$ (second per packet) for all the servers. The servers are randomly chosen to serve the packets by the scheduler, such that equal arrival rate is input to each server. We denote the utilization factor by $\rho = \lambda/N\mu$, and assume for stability that $\rho < 1$.

Note that if the server are statistically independent and the arrival process is independent of the service process, the scenario will be reduced to a classical M/M/N queueing network. Using queueing theory [2], the delay distribution and the mean delay of M/M/N model can be derived as follows;

$$\begin{aligned} Pr\{W(t) > t\} &= \frac{(N\rho)^N p_0}{N!(1-\rho)}e^{-N\mu(1-\rho)t} \\ \overline{W} &= \frac{(N\rho)^N p_0}{N\mu N!(1-\rho)^2} \end{aligned}, \tag{5}$$

where $p_0 = (\sum_{k=0}^{N-1}\frac{(N\rho)^k}{k!} + \frac{(N\rho)^N}{N!}\frac{1}{1-\rho})^{-1}$.

In what follows, we will use the stochastic network service curve to derive the delay bound distributions and the mean delay bounds for different cases. To derive the delay bound distribution, the following sufficient stability condition is widely used in the framework of stochastic network calculus (e.g. [8–12]).

$$\lim_{t\to+\infty}\frac{\alpha(t)}{t} \leq \lim_{t\to+\infty}\frac{\beta(t)}{t}, \tag{6}$$

Due to the fact that Poisson process has finite moment generating function, the SAC and the SSC with their bounding functions can be obtained through moment generating function in terms of [20]. The stochastic arrival curve with the bounding function for the multi-server network is

$$\begin{aligned} \alpha(t) &= \frac{1}{\theta}\log \mathrm{E}[e^{\theta A(t)}] = \frac{\lambda t}{\theta}(e^{\theta L} - 1) \\ f(x) &= e^{-\theta x} \end{aligned}. \tag{7}$$

In light of the system model description that the head packet of the buffer will be immediately served whenever there have idle servers, the arrival process for each server S_i $(1 \leq i \leq N)$ is also Poison process with mean rate $k_i\lambda$, where $k_i \in [0,1]$ denotes the proportion of the arrival traffic A being served by S_i and $\sum_{i=1}^{N} k_i = 1$. Hence the arrival curve with bounding

function for server S_i $(1 \leq i \leq N)$ holds

$$\begin{aligned} \alpha_i(t) &= \frac{k_i \lambda t}{\theta}(e^{\theta L} - 1) \\ f_i(x) &= e^{-\theta x} \end{aligned}. \tag{8}$$

The stochastic service curve with the bounding function for server $S_i (1 \leq i \leq N)$ is

$$\begin{aligned} \beta_i(t) &= -\frac{1}{\theta}\log \mathrm{E}[e^{-\theta S_i(t)}] = \frac{\mu t}{\theta}(1 - e^{-\theta L}) \\ g_i(x) &= e^{-\theta x} \end{aligned}. \tag{9}$$

According to Theorem 2 and Theorem 3, the stochastic network service curve holds as

$$\beta(t) = \sum_{i=1}^{N} \beta_i(t) = \frac{N\mu t}{\theta}(1 - e^{-\theta L}). \tag{10}$$

Applying the stability condition in (6), we have

$$\frac{\lambda}{\theta}(e^{\theta L} - 1) \leq \frac{N\mu}{\theta}(1 - e^{-\theta L}). \tag{11}$$

In Theorem 1, since both $f \in \bar{\Psi}$ and $g \in \bar{\Psi}$, it can easily be verified that the delay bound distribution also belongs to $\bar{\Psi}$. In order to minimize the delay bound, θ which conforms to (11) is optimized as

$$\theta = \frac{\ln(1/\rho)}{L}. \tag{12}$$

Furthermore, $h(\alpha + x, \beta)$ in Theorem 1 holds as

$$\beta(h(\alpha + x, \beta)) = x.$$

Replacing $h(\alpha + x, \beta)$ by t, there holds

$$x = \beta(t). \tag{13}$$

Delay bound analysis by applying the stochastic network service curve in Theorem 2:

At first, we derive the bounding function in Theorem 2 as follows

$$\begin{aligned} g(x) &= [f_1 \otimes \cdots \otimes f_N \otimes g_1 \otimes \cdots \otimes g_N(x - \sum_{i=1}^{N} \alpha_i \oslash \beta_i(0))]_1 \\ &= [2Ne^{-\frac{\theta x}{2N}}]_1 \end{aligned}. \tag{14}$$

We conduct our analysis for two cases, i.e. case 1: the arrival and service processes can either be correlated or independent; case 2: arrival and service processes are assumed to be exclusively independent.

Case 1: According to (1) in Theorem 1 and (13), the delay bound distribution holds as follows

$$Pr\{W(t) > t\} \leq [f \otimes g(\beta(t))]_1 \leq [(2N+1)e^{\frac{-N\mu(1-\rho)t}{2N+1}}]_1 \quad . \tag{15}$$

The mean delay bound holds as

$$\begin{aligned} \overline{W} &= \mathrm{E}[W(t)] = \int_0^{+\infty} Pr\{W(t) > t\}dt \\ &\leq \int_0^{+\infty} [(2N+1)e^{\frac{-N\mu(1-\rho)t}{2N+1}}]_1 dt \\ &= \int_0^{\frac{(2N+1)\ln(2N+1)}{N\mu(1-\rho)}} 1dt + \\ &\int_{\frac{(2N+1)\ln(2N+1)}{N\mu(1-\rho)}}^{+\infty} (2N+1)e^{\frac{-N\mu(1-\rho)t}{2N+1}}dt \\ &= \frac{(2N+1)\ln(2N+1) + (2N+1)}{N\mu(1-\rho)} \end{aligned} \quad . \tag{16}$$

Case 2: Similarly, we first derive the delay bound distribution by using (2) in Theorem 1. When $0 < t \leq \frac{2N\ln(2N)}{N\mu(1-\rho)}$, $1 - g(y) \equiv 0$, we have

$$Pr\{W(t) > t\} \leq 1 - \int_{-\infty}^{+\infty} (1 - f(x-y))d(1 - g(y)) = 1 \quad . \tag{17}$$

When $t > \frac{2N\ln(2N)}{N\mu(1-\rho)}$, according to (2) and (13), the delay bound distribution holds as follows

$$\begin{aligned} &Pr\{W(t) > t\} \\ \leq& 1 - \int_{-\infty}^{+\infty} (1 - f(x-y))d(1 - g(y)) \\ =& 1 - \\ &\int_{\frac{2N\ln(2N)}{N\mu(1-\rho)}}^{t} (1 - e^{-N\mu(1-\rho)(t-\tau)})d(1 - (N+1)e^{\frac{-N\mu(1-\rho)\tau}{N+1}}) \\ =& \frac{(2N)^2}{2N-1}e^{-\frac{N\mu(1-\rho)t}{2N}} - \frac{(2N)^{(2N)}}{2N-1}e^{-N\mu(1-\rho)t} \end{aligned} \quad . \tag{18}$$

The mean delay bound holds as

$$\begin{aligned} \overline{W} &\leq \int_0^{+\infty} Pr\{W(t) > t\}dt \\ &= \int_0^{\frac{2N\ln(2N)}{N\mu(1-\rho)}} 1dt + \int_{\frac{2N\ln(2N)}{N\mu(1-\rho)}}^{+\infty} (\frac{(2N)^2}{2N-1}e^{-\frac{N\mu(1-\rho)t}{2N}} \\ &\quad - \frac{(2N)^{(2N)}}{2N-1}e^{-N\mu(1-\rho)t})dt \\ &= \frac{2N\ln(2N) + (2N+1)}{N\mu(1-\rho)} \end{aligned} \quad . \tag{19}$$

Apparently, the stochastic network service curve in Theorem 3 will yield similar analysis when compared with Theorem 2, because the only difference between Theorem 2 and Theorem 3 is the bounding function $g(x)$. Hence, we omit the derivation for saving space and just give the results as depicted in TABLE I. In particular, the scenario in case 6 is actually an M/M/N queueing network. By the way, the approach in this subsection is applicable to other arrival and service processes as long as the SSC of each server is given or the service process for every server has finite moment generating function.

TABLE I
STOCHASTIC NETWORK CALCULUS RESULTS

Case	Delay Bound Distribution	Mean Delay Bound
Case 1: Theorem 2 arrival and service can either be independent or correlated servers can either be independent or correlated	$[(2N+1)e^{\frac{-N\mu(1-\rho)t}{2N+1}}]_1$	$\frac{(2N+1)\ln(2N+1)+(2N+1)}{N\mu(1-\rho)}$
Case 2: Theorem 2 arrival and service are independent servers can either be independent or correlated	$\begin{cases} 1, & t \le \frac{2N\ln(2N)}{N\mu(1-\rho)} \\ \frac{(2N)^2}{2N-1}e^{-\frac{N\mu(1-\rho)t}{2N}} - \frac{(2N)^{(2N)}}{2N-1}e^{-N\mu(1-\rho)t}, & t > \frac{2N\ln(2N)}{N\mu(1-\rho)} \end{cases}$	$\frac{2N\ln(2N)+(2N+1)}{N\mu(1-\rho)}$
Case 3: Theorem 3 arrival and service can either be independent or correlated servers can either be independent or correlated	$[(N+1)e^{\frac{-N\mu(1-\rho)t}{N+1}}]_1$	$\frac{(N+1)\ln(N+1)+(N+1)}{N\mu(1-\rho)}$
Case 4: Theorem 3 arrival and service are independent servers can either be independent or correlated	$\begin{cases} 1, & t \le \frac{N\ln N}{N\mu(1-\rho)} \\ \frac{N^2}{N-1}e^{-\frac{N\mu(1-\rho)t}{N}} - \frac{N^N}{N-1}e^{-N\mu(1-\rho)t}, & t > \frac{N\ln N}{N\mu(1-\rho)} \end{cases}$	$\frac{N\ln N+(N+1)}{N\mu(1-\rho)}$
Case 5: Theorem 3 arrival and service can either be independent or correlated servers are independent	$[2e^{\frac{-N\mu(1-\rho)t}{2}}]_1$	$\frac{2\ln 2+2}{N\mu(1-\rho)}$
Case 6 (**M/M/N model**): Theorem 3 arrival and service are independent servers are independent	$(1+N\mu(1-\rho)t)e^{-N\mu(1-\rho)t}$	$\frac{2}{N\mu(1-\rho)}$

IV. NUMERICAL RESULTS AND DISCUSSION

In this section, we present numerical results for the network scenario described in Section III-B. The total mean service rate is fixed as 5000 packets/s (i.e. $N\mu = 5000$ packet/s). The variables are the number of servers N and the service rate for each server μ. The results of analytical delay bounds for different cases will be compared with that of queueing theory. Also, the impacts of interdependence between the arrival and service processes and amongst the servers will be discussed.

A. Numerical Results

Fig. 2 shows the delay bound distributions under the assumption that the utilization factor $\rho = 0.8$, the number of servers $N = 10$ and the mean service rate for each server $\mu = 500$ packet/s. The probabilistic bounds derived based on Theorem 3 (i.e. cases 3-6) are tighter than the ones derived based on Theorem 2 (i.e. cases 1-2). This is because the results derived based on Theorem 2 dependents on both arrival and service processes while the ones based on Theorem 3 only dependents on service process. In particular, the analytical delay bound distribution for the M/M/N scenario (i.e., case 6) is close to the queueing theory result in (5), which confirms the accuracy of our analysis.

Fig.3 illustrates the mean queueing delay bounds under the assumption that number of servers $N = 10$ and the mean service rate for each server $\mu = 500$ packets/s. The mean delay bound increases with the utilization factor. Particularly, the increasing tendency is conspicuous when the input traffic becomes heavy (i.e. $\rho \ge 0.9$). Also, the mean delay bound derived by using Theorem 2 is more conservative than the one derived using Theorem 3. Furthermore, the subfig of Fig.3 shows that the analytical mean delay bound for the M/M/N scenario (i.e, case 6) is in accordance with the exact result in (5).

Fig. 4 depicts the relationship between the probabilistic delay violation bound and the number of servers. As the total

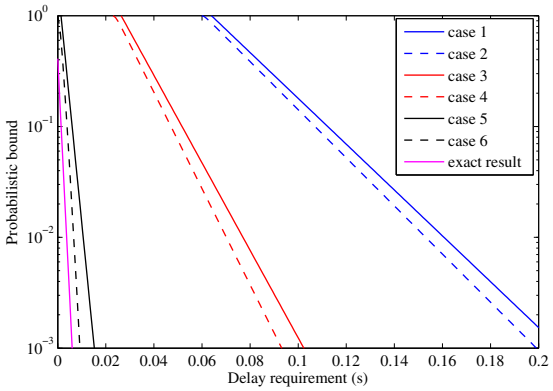

Fig. 2. Delay bound distribution-delay requirement curves

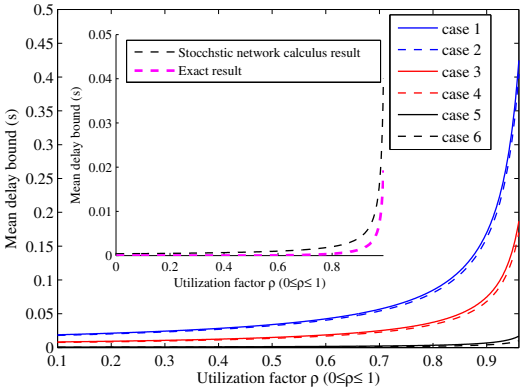

Fig. 3. Mean delay bound-utilization factor curves

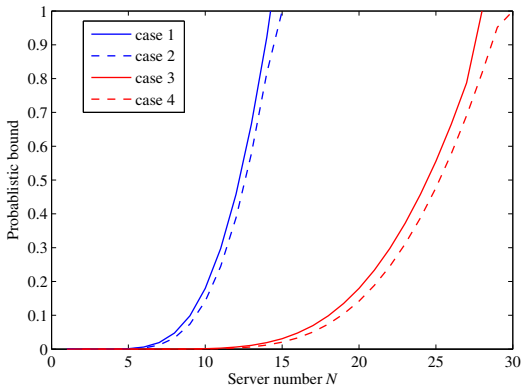

Fig. 4. Probabilistic delay violation bound-server number curves

mean service rate $N\mu$ is fixed, the number of servers N have no impact on the probabilistic bounds for cases 5 and 6 according to results in TABLE I. Consequently,we present results for cases 1-4 only where the servers can either be independent or correlated. The utilization factor ρ and delay requirement are set to 0.8 and 0.1s respectively. We observed that the probabilistic delay violation bound increases with N. Hence, Fig.4 suggests that it is more difficult for a multi-server network with fixed network capacity to guarantee the delay requirement whenever the network has more servers. In addition, we also observed that the probabilistic bounds derived by using Theorem 3 is tighter than the ones derived by using Theorem 2, which agrees with the results demonstrated in Fig. 2 and Fig. 3.

B. Accounting for Impacts of Interdependence

In what follows, we are going to discuss the impacts of interdependence between the arrival and service processes and amongst the servers.

At first, the six cases in TABLE I are divided into three subsets, which are subset 1 (including cases 1 and 2), subset 2 (including cases 3 and 4) and subset 3 (including cases 5 and 6). The only difference between the two cases in each subset is whether the arrival and service processes are independent or not. Fig. 2 and Fig .3 illustrate that in each subset, the independent case guarantees tighter probabilistic bound and smaller mean delay bound. Similarly, if we divided the six cases based on the same interdependence between the arrival and service processes, we get subset 1 (including cases 1, 3 and 5) and subset 2 (including cases 2, 4 and 6). In each subset, the delay performance for the case where servers are independent are better than the cases where servers can either be independent or correlated. Moreover, Fig. 2 and Fig. 3 also indicate that the impact of interdependence amongst the servers is much greater than that between the arrival and service processes.

V. Conclusion

In this paper, we use stochastic network calculus to perform delay analysis for a general multi-server network where the number of servers is arbitrary and the arrival and service processes are stochastic. General stochastic network service curves were derived in two ways and formulated as Theorem 2 and Theorem 3. Queueing delay bound distribution and mean queueing delay bound were derived for six cases (as classified in TABLE I) in a specific network. Moreover, our analytical results were proved to be close to the queuing theory results in an M/M/N network. We conclude that while analyzing a multi-server network, service processes which have finite moment generating functions can improve the analytical bounds. Also, a multi-server network could guarantee tighter delay bounds if the arrival and service processes are statistically independent or the servers are statistically independent. Moreover, the interdependence amongst the servers has greater impact on the analytical bounds than the one between the arrival and service processes.

References

[1] C. Y. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser ofdm with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct 1999.

[2] L. Kleinrock, "Queueing systems, volume i: Theory," 1975.

[3] P. Lu, Y. Yuan, Z. Yang, and Z. Zhu, "On the performance analysis of energy-efficient upstream scheduling for hybrid fiber-coaxial networks with channel bonding," *IEEE Commun. Lett.*, vol. 17, no. 5, pp. 1020–1023, 2013.

[4] H. Yang and B. Sikdar, "Queueing analysis of polling based wireless mac protocols with sleep-wake cycles," *IEEE Trans. Commun.*, vol. 60, no. 9, pp. 2427–2433, 2012.

[5] Y. Liu, M. Liu, and J. Deng, "Is diversity gain worth the pain: a delay comparison between opportunistic multi-channel mac and single-channel mac," in *Proc. IEEE INFOCOM*. IEEE, 2012, pp. 2921–2925.

[6] D. P. Bertsekas and R. G. Gallager, *Data networks*, vol. 2.

[7] C.-S. Chang, *Performance guarantees in communication networks*. Springer, 2000.

[8] Y. Jiang and Y. Liu, *Stochastic network calculus*. Springer, 2008.

[9] K. Zheng, F. Liu, L. Lei, C. Lin, and Y. Jiang, "Stochastic performance analysis of a wireless finite-state markov channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 782–793, February 2013.

[10] M. Beck, S. Henningsen, S. Birnbach, and J. Schmitt, "Towards a statistical network calculus dealing with uncertainty in arrivals," in *Proc. IEEE INFOCOM*, April 2014, pp. 2382–2390.

[11] K. Wu, Y. Jiang, and D. Marinakis, "A stochastic calculus for network systems with renewable energy sources," in *IEEE INFOCOM Workshops*, March 2012, pp. 109–114.

[12] Z. Li, Y. Gao, L. Sang, and D. Yang, "Analysis on the energy consumption in stochastic wireless networks," in *IEEE ICC Workshops*, June 2014, pp. 866–870.

[13] Y. Jiang, "Network calculus and queueing theory: two sides of one coin: invited paper," in *Proc. 4th Int. ICST Conf. Performance Evaluation Methodologies and Tools*. ICST, 2009, p. 37.

[14] F. Ciucu, "Network calculus delay bounds in queueing networks with exact solutions," in *Managing Traffic Performance in Converged Networks*. Springer, 2007, pp. 495–506.

[15] Y. Gao, W. Jiang, and Y. Jiang, "Guaranteed service and delay-constrained capacity of a multi-channel cognitive secondary network," in *Proc. IEEE CROWNCOM*. IEEE, 2012, pp. 83–88.

[16] X. Haiming and Y. Jiang, "Analysis of multi-server round robin scheduling disciplines," *IEICE trans. commun.*, vol. 87, no. 12, pp. 3593–3602, 2004.

[17] S. Kittipiyakul and T. Javidi, "Delay-optimal server allocation in multiqueue multiserver systems with time-varying connectivities," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2319–2333, May 2009.

[18] S. Bodas, S. Shakkottai, L. Ying, and R. Srikant, "Scheduling in multi-channel wireless networks: Rate function optimality in the small-buffer regime," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1101–1125, Feb 2014.

[19] ——, "Low-complexity scheduling algorithms for multichannel downlink wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1608–1621, Oct 2012.

[20] Y. Jiang, "A note on applying stochastic network calculus. http://q2s.ntnu.no/ jiang/publications.html," 2010.