Network Calculus and Queueing Theory: Two Sides of One Coin

[Invited Paper]

Yuming Jiang

Centre for Quantifiable Quality of Service in Communication Systems *
Department of Telematics
Norwegian University of Science and Technology (NTNU)

ABSTRACT

Network calculus is a theory dealing with queueing type problems encountered in computer networks, with particular focus on quality of service guarantee analysis. Queueing theory is the mathematical study of queues, proven to be applicable to a wide area of problems, generally concerning about the (average) quantities in an equilibrium state. Since both network calculus and queueing theory are analytical tools for studying queues, a question arises naturally as is if and where network calculus and queueing theory meet. In this paper, we explore queueing principles that underlie network calculus and exemplify their use. Particularly, based on the network calculus queueing principles, we show that for $GI/GI/1$, similar inequalities in the theory of queues can be derived. In addition, we prove that the end-to-end performance of a tandem network is independent of the order of servers in the network even under some general settings. Through these, we present a network calculus perspective on queues and relate network calculus to queueing theory.

1. INTRODUCTION

Queueing theory is the general mathematical study of queues. In 1909, Danish mathematician and engineer A. K. Erlang [2] published "The Theory of Probabilities and Telephone Conversations" that originated the field of queueing theory. Since then, queueing theory has been developed and applied in a wide variety of areas including engineering, business and public service. The *classical queueing theory*, or *queueing theory* in short, generally concerns about the (average) quantities in an equilibrium state. It has played a fundamental role in modeling, analyzing and dimensioning communication networks. For example, the very first queueing theory paper proves that the Poisson distribution applies to random telephone traffic.

With the advance of communication and networking technologies, it is natural to apply queueing theory also to modern packet-switched computer networks such as the Internet. However, unique customer and service characteristics and requirements in such networks often make the application difficult. Due to this, there has been a demand for a new theory to deal with queueing type problems encountered in computer networks. Since early 1990s, the development of this new theory has attracted a lot of research attention and effort, and significant progress has been made. The theory is now known under the name of *network calculus*.

Network calculus is a theory dealing with queueing type problems encountered in modern packet-switched computer networks. Its focus is on performance guarantees. Central to the theory is the use of alternate algebras such as min-plus algebra and max-plus algebra to transform complex network systems into analytically tractable systems. To simplify the analysis, another idea is to characterize the arrival and service processes using some bounds and base performance analysis on such bounds. Network calculus has developed along two tracks — deterministic and stochastic.

The idea of using a function to deterministically upper-bound the cumulative arrival process was initially introduced by Cruz in the seminal work [10]. This idea results in the arrival curve model in network calculus for arrival modeling. For server modeling, a similar idea was introduced in [25], which uses a function to deterministically lower-bound the cumulative service process. This idea has evolved into the service curve model in network calculus. Also for server modeling, there is another promising idea, which is to compare the actual departure time with a virtual time function, and use the difference together with the rate parameter, which is the basis of the virtual time function, to model the actual server. This idea was initially used in [30] to define the Virtual Clock scheduling algorithm. In [12], the idea was explored to define a server model, called guaranteed rate (GR) server. Evolved from these models, a lot of results have been derived for deterministic network calculus and excellent books are available [6][19].

The development of stochastic network calculus began also in early 1990s. Early representative works include [18] [29] [4] for arrival modeling, and [20] for server modeling. Essentially, the arrival models and server models of stochastic network calculus can be considered as probabilistic extension or indeed generalization of their counterparts in determin-

*The "Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" is appointed by The Research Council of Norway, and funded by The Research Council, NTNU and UNINETT. http://www.q2s.ntnu.no

VALUETOOLS 2009, October 20-22, 2009 - Pisa, Italy.

istic network calculus. However, due to challenges specific to stochastic networks, it is recently that crucial network calculus properties have been proved for stochastic network calculus. Representative works include [3][9][21][13][23][16] and [11]. A book summarizing the development and results is available [15].

Since both network calculus and queueing theory are analytical tools for studying queues, it is natural to ask if and where they meet. Particularly, in order to relate network calculus to queueing theory, we need to answer the following questions. Are there some queueing principles underlying network calculus? What are they? Can the network calculus queueing principles be explored to study classical queueing problems, and how?

The intention of this paper is to take one step forward towards providing answers / insights to the above questions. In particular, through reviewing the fundamental concepts and models of network calculus, we explore queueing principles that underlie network calculus. In addition, we exemplify their use in tackling some classical queueing problems. We show that for $GI/GI/1$, similar inequalities in the theory of queues can be derived based on the network calculus queueing principles. Also, we prove that the end-to-end performance of a tandem network is independent of the order of servers in the network even under some general settings. Through these, we present a network calculus perspective on queues and relate network calculus to queueing theory.

The rest is organized as follows. After introducing the system model and Lindley recursion in the next section, we present fundamental concepts and models of network calculus in Section 3. In Section 4, we discuss and expose two network calculus queueing principles, based on which, analysis on single server queue and tandem network is performed in Section 5 and Section 6 respectively. Finally, the paper is summarized in Section 7.

2. PRELIMINARIES

2.1 System Model and Notation

In this paper, we consider single-server queue systems. Each queue is assumed to be first-in-first-out (FIFO) with infinite queue space. All queues are empty at time 0. For such a system, we define notation in the following. A summary of the notation is provided in Table 1.

We use $C(n)$ to denote the $(n+1)$st customer to enter the system, where $n = 0, 1, \ldots$. Its arrival time to the system is $a(n)$ and departure time from the system is $d(n)$. The service time of $C(n)$ is $\delta(n)$. The inter-arrival time between $C(n)$ and $C(n+1)$ is $\tau(n)$. The inter-arrival time between $C(m)$ and $C(n)$ is denoted by $\Gamma(m,n) = \sum_{k=m}^{n-1} \tau(k)$. The cumulative service time of customers $C(m)$ to $C(n)$ is $\Delta(m,n) = \sum_{k=m}^{n} \delta(k)$. By definition, the delay of $C(n)$, denoted by $D(n)$, is:

$$D(n) = d(n) - a(n). \tag{1}$$

In this paper, we shall also adopt another way to model the system. We use $A(t)$ to denote the amount of required service from customers entering the system up to time t, $S(t)$ the amount of provided service, up to time t, by the system to the arrivals, and $A^*(t)$ the output in terms of required service amount from the system up to time t. By convention, we adopt $A(0) = S(0) = A^*(0) = 0$. In addition, we let $A(s,t) \equiv A(t) - A(s)$, $S(s,t) \equiv S(t) - S(s)$, and $A^*(s,t) \equiv A^*(t) - A^*(s)$. Under this model, the delay at time t is defined as

$$D(t) = \inf\{d \geq 0 : A(t) \leq A^*(t+d)\} \tag{2}$$

and the backlog at time t is defined as

$$B(t) = A(t) - A^*(t). \tag{3}$$

Table 1: Notation

$C(n)$	The $(n+1)$st customer, $n = 0, 1, 2, \ldots$
$a(n)$	Arrival time of $C(n)$
$A(t)$	Cumulative arrival up to time t
$A(s,t)$	Cumulative arrival in $(s,t]$: $A(s,t) = A(t) - A(s)$
$d(n)$	Departure time of $C(n)$
$A^*(t)$	Cumulative departure up to time t
$\tau(n)$	Inter-arrival time between $C(n)$ and $C(n+1)$
$\Gamma(m,n)$	Inter-arrival time between $C(m)$ and $C(n)$: $\Gamma(m,n) = \sum_{k=m}^{n-1} \tau(k)$
$\delta(n)$	Service time of $C(n)$
$\Delta(m,n)$	Cumulative service time of $C(m)$ to $C(n)$: $\Delta(m,n) = \sum_{k=m}^{n} \delta(k)$
$S(t)$	Amount of provided service up to time t
$S(s,t)$	Cumulative amount of service in $(s,t]$: $S(s,t) = S(t) - S(s)$
$u(n)$	$\equiv \delta(n) - \tau(n)$
$D(n)$	System delay (queue plus service) of $C(n)$
$D(t)$	Delay at time t: $D(t) = \inf\{d \geq 0 : A(t) \leq A^*(t+d)\}$
D	$\equiv \lim_{n\to\infty} D(n)$: Expected system delay
$W(n)$	Waiting time in queue of $C(n)$
W	$\equiv \lim_{n\to\infty} W(n)$
$B(t)$	Backlog in the system at time t: $B(t) = A(t) - A^*(t)$

2.2 Stochastic Ordering

For any two random variables X and Y, if $P\{X > x\} \leq P\{Y > x\}$ for all x, we say X is stochastically smaller than Y [26], written as:

$$X \leq_{st} Y.$$

The same notation applies when X and Y are random vectors, and the following result holds (e.g. see (1.10.5) in [26]).

LEMMA 1. *Let $X = \{X_1, \ldots, X_N\}$ and $Y = \{Y_1, \ldots, Y_N\}$ be N-vectors of random variables. If $X_1, \ldots, X_N$ are mutually independent and so are $Y_1, \ldots, Y_N$, then $X \leq_{st} Y$, or in other words $\{X_1, \ldots, X_N\} \leq_{st} \{Y_1, \ldots, Y_N\}$, if and only if $X_i \leq_{st} Y_i$, $(i = 1, \ldots, N)$.*

Consider a function or mapping as follows:

$$Y = \Phi(X_1, \ldots, X_N) \tag{4}$$

where $X_1, \ldots, X_N$ are random variables, Φ is the function or mapping and Y is the resulting random variable. For such a mapping, the following result holds (e.g. see Theorem 2.2.4 in [26]).

LEMMA 2. *Let $Y = f(X_1, \ldots, X_n, X_{n+1}, \ldots, X_N)$ and $Y' = \Phi(X_1, \ldots, X_n, X'_{n+1}, \ldots, X'_N)$. Suppose the random variables $\{X_{n+1}, \ldots, X_N\}$ and $\{X'_{n+1}, \ldots, X'_N\}$ are independent of $\{X_1, \ldots, X_n\}$. When Φ is nondecreasing in $\{x_{n+1}, \ldots, x_N\}$,*

then, $\{X_{n+1}, \ldots, X_N\} \leq_{st} \{X'_{n+1}, \ldots, X'_N\}$ *implies* $Y \leq_{st} Y'$.

Lemma 1 and Lemma 2 will be used heavily in proving Theorem 3 and Theorem 4 later in this paper.

2.3 Lindley Recursion

In the literature, Lindley recursion is (perhaps the most) widely used queueing principle in analyzing single server queue systems, which is:

$$W(n) = [W(n-1) + u(n-1)]^+ \tag{5}$$

where $u(n-1) = \delta(n-1) - \tau(n-1)$. Note that equation (5) holds without any requirement on the arrival process or the service process, and is a principle of queueing. In this paper, we shall call recursion (5) the "*Lindley recursion queueing principle*".

Based on the Lindley recursion queueing principle, Kingman proved the following inequality for $GI/GI/1$ [17]:

$$P\{W \geq x\} \leq e^{-\theta_0 x} \tag{6}$$

where θ_0 is found from

$$\theta_0 = \sup\{\theta > 0 : M_{\delta(0)-\tau(0)}(\theta) < 1\}$$

and $M_X(\theta)$ denotes the moment generating function of X:

$$M_X(\theta) \equiv E[e^{\theta X}].$$

In this paper, we shall show that similar inequalities can be proved based on network calculus queueing principles.

3. NETWORK CALCULUS BASICS

An essential idea of network calculus is to use min-plus algebra and max-plus algebra to transform non-linear queueing systems into linear systems that are analytically tractable.

In min-plus algebra, the algebra structure of interest is $(\mathcal{R} \cup \{+\infty\}, \wedge, +)$. Here, the "addition" operation is $\wedge$ and the "multiplication" operation is $+$, where $\wedge$ denotes the *infimum* or, when it exists, the *minimum*. Similarly, in max-plus algebra, the algebra structure of interest is $(\mathcal{R} \cup \{+\infty\}, \vee, +)$. Here, the "addition" operation is $\vee$ and the "multiplication" operation is $+$, where $\vee$ denotes the *supremum* or, when it exists, the *maximum*. It can be verified that both $(\mathcal{R} \cup \{+\infty\}, \wedge, +)$ and $(\mathcal{R} \cup \{+\infty\}, \vee, +)$ have similar properties as the conventional algebra such as the closure property, associativity, commutativity, and distributivity.

As in the conventional algebra, min-plus algebra and max-plus algebra also have their respective convolution operation. The min-plus convolution, denoted by $\otimes$, and the max-plus convolution, denoted by $\bar{\otimes}$, are respectively defined as:

$$F \otimes G(t) = \inf_{0 \leq \tau \leq t} \{F(\tau) + G(t-\tau)\}, \tag{7}$$

and

$$a \bar{\otimes} b(n) = \sup_{0 \leq k \leq n} \{a(k) + b(n-k)\}. \tag{8}$$

It can be verified that $(\mathcal{F}, \wedge, \otimes)$ forms a complete dioid, where $\mathcal{F}$ denotes the set of nonnegative nondecreasing single variate functions. Similarly, it can be verified that $(\mathcal{F}, \vee, \bar{\otimes})$ also forms a complete dioid.

To help exposition, we specifically list the associativity of min-plus and max-plus convolutions, which is:

- (Associativity of min-plus convolution) $\forall F, G, H \in \mathcal{F}$

$$(F \otimes G) \otimes H = F \otimes (G \otimes H)$$

- (Associativity of max-plus convolution) $\forall a, b, c \in \mathcal{F}$

$$(a \bar{\otimes} b) \bar{\otimes} c = a \bar{\otimes} (b \bar{\otimes} c)$$

In addition, the commutativity property of convolution is:

- (Commutativity of min-plus convolution) $\forall F, G \in \mathcal{F}$

$$F \otimes G(t) = G \otimes F(t)$$

- (Commutativity of max-plus convolution) $\forall a, b \in \mathcal{F}$

$$a \bar{\otimes} b(n) = b \bar{\otimes} a(n).$$

3.1 Deterministic Network Calculus

Arrival curve and service curve are the most fundamental concepts and models for deterministic network calculus. While an arrival curve defines a bound on the arrival, a service curve defines a bound on the service. There are two basic ways to define such a bound. One is to bound the total amount of required service of the arrivals or the total amount of provided service in a time period; another is to bound the inter-arrival or inter-service time. We shall call the former *space domain* modeling, and the latter *time domain* modeling.

Under space domain modeling, an input is said to have an arrival curve $\alpha(t) \in \mathcal{F}$, if the following inequality holds for all $t \geq 0$:

$$A(t) \leq A \otimes \alpha(t). \tag{9}$$

A server is said to provide a service curve $\beta(t) \in \mathcal{F}$, if there holds for all $t \geq 0$:

$$A^*(t) \geq A \otimes \beta(t). \tag{10}$$

Essentially, the above arrive curve and service curve models are defined based on cumulative arrivals and cumulative service. We shall call them in the rest of the paper the min-plus arrival curve and the min-plus service curve models respectively. Based on the min-plus arrival curve and service curve models, a lot of results have been derived and excellent books summarizing them are available [19][6].

Among the others, one important result of network calculus, which is called the *concatenation property*, is that, for a network of H tandem servers, if each server h provides to its input a service curve $\beta_h(t)$, then the whole network provides to the input a (network) service curve β as:

$$\beta(t) = \beta_1 \otimes \cdots \otimes \beta_H(t). \tag{11}$$

Under time domain modeling, there is a similar pair of arrival and service curve models, which we shall call the max-plus arrival curve and the max-plus service curve respectively. They are defined as follows.

An input is said to have a max-plus arrival curve $\gamma(n) \in \mathcal{F}$, if the following inequality holds for all $n \geq 0$:

$$a(n) \geq a \bar{\otimes} \gamma(n). \tag{12}$$

A server is said to provide a max-plus service curve $\eta(n) \in \mathcal{F}$, if there holds for all $n \geq 0$:

$$d(n) \leq a \bar{\otimes} \eta(n). \tag{13}$$

Based on the time-domain max-plus arrival curve and service curve models, similar results for network service guarantee analysis have been derived (e.g. see [12][7]). For the network of tandem servers, if each server h provides to its input a max-plus service curve $\eta_h(n)$, then the whole network provides to the input a (network) max-plus service curve η as:

$$\eta(n) = \eta_1 \bar{\otimes} \cdots \bar{\otimes} \eta_H(n). \tag{14}$$

We notice that the space domain min-plus arrival curve and service curve models and results based on them are probably more widely known [6][19] than the time-domain models and results. Nevertheless, we would like to stress that they compliment each other. For example, the max-plus service curve has special properties that have been explored in service guarantee analysis and provisioning. An in-depth discussion is out of the scope of this paper and readers may refer to [14] for some discussion.

Owing to the associativity and commutativity of min-plus convolution and max-plus convolution, the network service curve under both min-plus network calculus and max-plus calculus for the tandem network case does not change no matter how servers in the network are ordered. Since the end-to-end performance guarantee for such a network is decided by the network service curve and the input arrival curve, we can conclude that changing the order of servers in the tandem does not affect the end-to-end performance guarantee analysis results.

3.2 Stochastic Network Calculus

Stochastic network calculus is the probabilistic extension or indeed generalization of deterministic network calculus. It provides a more natural link to queueing theory, since deterministic network calculus has its basis on worst-case analysis avoiding the stochastic nature of queueing processes.

For simplicity, we shall only introduce the direct generalization of the deterministic arrival and server models reviewed in the previous subsection. Readers may refer to [15][28] for various variations of these models.

Denote by $\bar{\mathcal{F}}$ the set of nonnegative nonincreasing functions. The min-plus arrival curve and service curve models can be generalized as:

An input is said to have a min-plus stochastic arrival curve $\alpha(t) \in \mathcal{F}$ with bounding function $f \in \bar{\mathcal{F}}$, if for any $x \geq 0$, the following holds for all $t \geq 0$:

$$P\{A(t) - A \otimes \alpha(t) > x\} \leq f(x). \tag{15}$$

A server is said to provide a (min-plus) stochastic service curve $\beta(t) \in \mathcal{F}$ with bounding function $g \in \bar{\mathcal{F}}$, if there holds for all $t \geq 0$:

$$P\{A \otimes \beta(t) - A^*(t) > x\} \leq g(x). \tag{16}$$

For the stochastic case, the *concatenation property* has also been proved. Particularly, for a network of H tandem servers, if each server h provides to its input a stochastic service curve $\beta_h(t)$ with bounding function $g_h \in \bar{\mathcal{F}}$, then the whole network provides to the input a (network) stochastic service curve β as [15]:

$$\beta(t) = \beta_1 \otimes \beta_2^{-\theta} \cdots \otimes \beta_H^{-(H-1)\theta}(t) \tag{17}$$

with bounding function g as

$$g(x) = g_1^{\theta_1} \otimes g_2^{\theta_2} \cdots \otimes g_H^{\theta_H}(x) \tag{18}$$

for any $\theta, \theta_1, \ldots, \theta_H > 0$ where $\beta_h^{(h-1)\theta}(t) = [\beta_h(t) - (h-1)\theta t]^+$ and $g_h^{\theta_h}(x) = g_h(x) + \frac{1}{\theta_h}\int_x^\infty g_h(y)dy$, $h = 1, \ldots, H$.

Similarly, the max-plus arrival curve and service curve models can be generalized as [28]:

An input is said to have a max-plus stochastic arrival curve $\gamma(n) \in \mathcal{F}$, with bounding function $f \in \bar{\mathcal{F}}$, if for any $x \geq 0$, the following holds for all $n \geq 0$:

$$P\{a \bar{\otimes} \gamma(n) - a(n) > x\} \leq f(x). \tag{19}$$

A server is said to provide a max-plus stochastic service curve $\eta(n) \in \mathcal{F}$, with bounding function $g \in \bar{\mathcal{F}}$, if for any $\delta \geq 0$, the following holds for all $n \geq 0$:

$$P\{d(n) - a \bar{\otimes} \eta(n) > x\} \leq g(x). \tag{20}$$

For the stochastic case, similar *concatenation property* can be proved. Particularly, for a network of H tandem servers, if each server h provides to its input a stochastic service curve $\eta(n)$ with bounding function $g_h \in \bar{\mathcal{F}}$, then the whole network provides to the input a (network) stochastic service curve η as

$$\eta(n) = \eta_1 \bar{\otimes} \eta_2^{\theta} \cdots \otimes \eta_H^{(H-1)\theta}(n) \tag{21}$$

with bounding function g as

$$g(x) = g_1^{\theta_1} \otimes g_2^{\theta_2} \cdots \otimes g_H^{\theta_H}(x) \tag{22}$$

for any $\theta, \theta_1, \ldots, \theta_H > 0$ where $\eta_h^{(h-1)\theta}(n) = \eta_h(n) + (h-1)\theta n$ and $g_h^{\theta_h}(x) = g_h(x) + \frac{1}{\theta_h}\int_x^\infty g_h(y)dy$, $h = 1, \ldots, H$.

Similar to the deterministic case, also owing to the associativity and commutativity of min-plus convolution and max-plus convolution, and the arbitrariness of $\theta, \theta_1, \ldots, \theta_H$ in (17) to (22), changing the order of servers in the tandem typically does not affect the end-to-end stochastic service curve[1] or its bounding function, nor does the end-to-end stochastic service guarantee analysis results.

4. NETWORK CALCULUS QUEUEING PRINCIPLES

Having introduced the fundamental models of network calculus, we expose in this section two basic queueing principles that underlie these models.

4.1 Min-Plus Convolution Queueing Principle

Under discrete time, the Lindley recursion has another form that is:

$$B(t) = [B(t-1) + A(t-1,t) - S(t-1,t)]^+ \tag{23}$$

which, when applied iteratively, results in

$$B(t) = \sup_{0 \leq s \leq t} [A(s,t) - S(s,t)]. \tag{24}$$

Since there holds $A^*(t) = A(t) - B(t)$ by definition, we then have

$$A^*(t) = \inf_{0 \leq s \leq t} [A(s) + S(s,t)]. \tag{25}$$

With a bit abuse of notation, we define min-plus convolution of two bivariate functions F and G as:

$$F \otimes G(s,t) = \inf_{s \leq \tau \leq t} \{F(s,\tau) + G(\tau,t)\}. \tag{26}$$

[1]Strictly speaking, there can be some effect on the resultant network service curve after changing the order of servers.

Then, equation (25) can be re-written as:

$$A^*(0,t) = A \otimes S(0,t). \quad (27)$$

Equations (27) and (10) resemble each other. Comparing (27) and (10), it is clear that if for any $0 \leq s \leq t$, the service $S(s,t)$ in this period, which is random in nature, is *lower-bounded* by a function $\beta(t-s)$, then the system has a service curve β. In other words, equation (27) can be considered as the queueing basis of the various min-plus service curve models and consequently the space-domain min-plus network calculus.

It is worth highlighting that equation (25) and equivalently (27) hold in general as the Lindley recursion queueing principle does. We shall hence call equation (25) and (27) the "*min-plus convolution queueing principle*". It can be verified that (25) is linear under min-plus algebra [15].

4.2 Max-Plus Convolution Queueing Principle

Like the Lindley recursion queueing principle and the min-plus convolution queueing principle, there is another recursion that holds in general in queueing systems, which corresponds to the time-domain service curve models.

Consider the departure time of customer $C(n)$. If this customer arrives to the system after the previous customer $C(n-1)$ having finished the service, then the departure time of $C(n)$ is simply $a(n)+\delta(n)$. However, if customer $C(n)$ arrives seeing customer $C(n-1)$ still in the system, then its departure time will be $d(n-1)+\delta(n)$. Combining both cases, we have the following:

$$d(n) = \max\{a(n), d(n-1)\} + \delta(n). \quad (28)$$

Applying (28) iteratively to its right hand side results in:

$$d(n) = \max_{0\leq m\leq n}\{a(m)+\Delta(m,n)\} \quad (29)$$

Again with a bit abuse of notation, we define max-plus convolution of two bivariate functions a and b as

$$a\bar{\otimes}b(m,n) = \sup_{m\leq k\leq n}\{a(m,k)+b(k,n)\}. \quad (30)$$

Then, equation (28) can be written as

$$d(n) = a\bar{\otimes}\Delta(0,n), \quad (31)$$

where we adopt $a(0,k) \equiv a(k)$.

Equations (31) and (13) also resemble each other. Comparing them, it is clear that if for the cumulative service time $\Delta(m,n)$, which is random in nature, is *upper-bounded* by a function $\eta(n-m)$, then the system has a max-plus service curve η. Implicitly, equation (31) provides the queueing basis and lays the foundation for max-plus service curve models and consequently the time domain max-plus network calculus.

As it is clear in the discussion, (28) holds in general. We shall hence call recursion (28) and equation (31) the "*max-plus convolution queueing principle*".

It is noticed that the queueing theory literature has a lot of results based on the Lindley queueing principle and so has the network calculus literature based on the min-plus convolution principle. However, there are relatively fewer results based on the max-plus convolution principle in the literature. Nevertheless, the max-plus convolution principle provides another perspective on viewing queueing problems. In this paper, we shall also explore this new perspective and elaborate its use.

4.3 Discussion

While the three discussed queueing principles do not resemble each other in expression, they are indeed closely related.

In the introduction to the min-plus convolution queueing principle, we have applied the Lindley recursion principle. In addition, the following discussion shows that the max-plus convolution queueing principle is also related to the Lindley recursion principle.

It is trivial that the system delay of customer $C(n)$ is $D(n) = d(n) - a(n)$, applying which to the max-plus convolution principle or equation (29) results in:

$$\begin{aligned} D(n) &= \max_{1\leq m\leq n}\{\Delta(m,n) - \Gamma(m,n)\} \\ &= \max_{0\leq m\leq n}\left[\sum_{k=m}^{n}\delta(k) - \sum_{k=m}^{n-1}\tau(k)\right]. \end{aligned} \quad (32)$$

Since the waiting time in queue of $C(n)$ is $W(n) = D(n) - \delta(n)$, we then have

$$W(n) = \max_{0\leq m\leq n}\left[\sum_{k=m}^{n-1}\delta(k) - \sum_{k=m}^{n-1}\tau(k)\right]. \quad (33)$$

It can be verified that iteratively applying (5) to its right hand side results in the same expression as the right hand side of (33). In other words, the Lindley principle can be derived from the max-plus queueing convolution principle and vice versa, which further provides the link between the two network calculus queueing principles.

Nevertheless, we stress that the two network calculus queueing principles have special properties that may not be found from the Lindley principle. Particularly, for the network case, it is difficult to find an extension of the Lindley recursion. On the other hand, as to be introduced in Sec. 6, both the min-plus convolution queueing principle and the max-plus convolution queueing principle can be easily extended to the network case, which can be explored to study queueing problems that could otherwise be difficult with the Lindley principle.

5. SINGLE SERVER QUEUES

In this section, we demonstrate how to make use of the min-plus and max-plus queueing principles to analyze single server queues. The focus is on establishing bounds on the tail of delay distribution in $GI/GI/1$.

In $GI/GI/1$, the arrivals are independent and identically distributed (i.i.d.), so are the service times. Then, the Lindley recursion implies that the waiting times of customers form a Markov chain. In addition, there holds [26]:

$$W(n) \leq_{st} W(n+1). \quad (34)$$

This monotonicity property tells that the equilibrium waiting time distribution is an upper bound.

5.1 Min-Plus Queueing Principle Approach

For ease of exposition in this subsection, we assume discrete time with unit discretization step [2].

The delay definition implies (e.g. see [15]): for any $x \geq 0$,

$$\{D(t) > x\} \subset \{A(t) > A^*(t+x)\}.$$

[2]The continuous time case is approximated when the time length of the unit approaches 0.

Based on the min-plus queueing principle, we get

$$A(t) - A^*(t+x) = \sup_{0\le s\le t+x}\{A(t) - A(s) - S(s, t+x)\},$$

and then

$$\begin{aligned} & P\{A(t) > A^*(t+x)\} \\ = \; & P\{\sup_{0\le s\le t-1}\{A(s,t) - S(s,t+x)\} > 0\} \\ = \; & P\{\sup_{0\le s\le t-1} e^{A(s,t)-S(s,t+x)} > 1\}. \end{aligned} \quad (35)$$

Note that the assumption of i.i.d. arrivals and i.i.d. service times implies $A(t)$ has independent and stationary increments and so has $S(t)$. In addition, we assume there exists some $\theta > 0$ making $M_{A(1)-S(1)}(\theta) < 1$. Let $\theta_0 = \sup\{\theta : M_{A(1)-S(1)}(\theta) < 1\}$. Based on these, in the following, we present two bounds on delay, which have the same form as (6).

Applying first Boole's inequality to the right hand side of (35) and then the Chernoff bound, we obtain:

$$\begin{aligned} & P\{\sup_{0\le s\le t-1} e^{A(s,t)-S(s,t+x)} > 1\} \\ \le \; & \sum_{s=0}^{t-1} P\{e^{A(s,t)-S(s,t+x)} > 1\} \\ \le \; & e^{-\theta}\sum_{s=0}^{t-1} E[e^{\theta(A(s,t)-S(s,t+x))}] \\ \le \; & \frac{M_{A(1)-S(1)}(\theta)}{1 - M_{A(1)-S(1)}(\theta)} e^{-\theta}\left[M_{S(1)}(-\theta)\right]^x \end{aligned} \quad (36)$$

where and in the rest of the paper, we adopt $M_X(-\theta) \equiv E[e^{-\theta x}]$ for random variable X. So, we have the following bound on delay:

$$\begin{aligned} & P\{D(t) > x\} \\ \le \; & \inf_{0<\theta\le\theta_0}\left\{\frac{M_{A(1)-S(1)}(\theta)}{1 - M_{A(1)-S(1)}(\theta)} e^{-\theta}\left[M_{S(1)}(-\theta)\right]^x\right\} \end{aligned} \quad (37)$$

For the other bound, it is based on the property of martingale. Note that the right hand side of (35) can be written as

$$\begin{aligned} & P\{\sup_{0\le s\le t-1} e^{A(s,t)-S(s,t+x)} > 1\} \\ = \; & P\{\sup_{1\le u\le t} e^{A(t-u,t)-S(t-u,t+x)} > 1\}. \end{aligned} \quad (38)$$

Fixing t, consider the random process as

$$V(u) \equiv e^{\theta[A(t-u,t)-S(t-u,t+x)]}$$

with $1 \le u \le t$ and $\theta > 0$, which is associated with $\sigma-$algebra $\mathcal{M}_u$. Note that $A(t-u-1, t-u) - S(t-u-1, t-u)$ is independent of $A(t-v,t) - S(t-v, t+x)$, $v = 1, \ldots, u$. If $M_{A(1)-S(1)}(\theta) \le 1$, we then have

$$\begin{aligned} & E[V(u+1)|\mathcal{M}_u] \\ = \; & E[V(u)e^{\theta[A(t-u-1,t-u)-S(t-u-1,t-u)]}|\mathcal{M}_u] \\ = \; & V(u)E[e^{\theta[A(t-u-1,t-u)-S(t-u-1,t-u)]}] \\ = \; & V(u)E[e^{\theta[A(1)-S(1)]}] \le V(u) \end{aligned} \quad (39)$$

which indicates that $V(u)$, $(u = 1, \cdots, t)$ is a supermartingale. Then, based on Doob's inequality for martingale, we obtain the following delay bound:

$$\begin{aligned} & P\{D(t) > x\} \\ \le \; & P\{A(t) > A^*(t+x)\} = P\{\sup_{1\le u\le t} V(u) > 1\} \\ \le \; & E[V(1)] = E[e^{\theta[A(t-1,t)-S(t-1,t+x)]}] \\ = \; & M_{A(1)-S(1)}(\theta)\left[M_{S(1)}(-\theta)\right]^x. \end{aligned} \quad (40)$$

So far, we have proved two bounds for $D(t)$. Both have similar form as the Kingman's bound. The bound based on supermartingale is generally tighter than the Chernoff bound, since θ is normally small. In addition, for any t, there is $a(n-1) \le t < a(n)$ where $n = \inf\{m : A(a(m)) > A(t)\}$, based on which, the following relationship can be verified:

$$[d(n-1) - a(n)]^+ < D(t) \le d(n-1) - a(n-1). \quad (41)$$

Note that the first term in above is indeed the waiting time of $C(n)$, i.e. $[d(n-1) - a(n)]^+ = W(n)$. We have hence derived bounds for waiting time in queue and proved the following theorem.

THEOREM 1. *Consider a $GI/GI/1$ queue. We assume $M_{A(1)-S(1)}(\theta)$ exists for small $\theta > 0$. Let $\theta_0 = \sup\{\theta : M_{A(1)-S(1)}(\theta) \le 1\}$. Then, we have the following bound for waiting time in queue:*

$$P\{W > x\} \le \inf_{0<\theta\le\theta_0}\left\{M_{A(1)-S(1)}(\theta)\left[M_{S(1)}(-\theta)\right]^x\right\}. \quad (42)$$

Remark: For the $GI/GI/1$ queue, we shall interpret $A(t)$ as the *number of customers* that have arrived up to time t. In other words, the unit of required service is the customer. Note that $A(t)$ may be interpreted differently when the unit of required service is different. Particularly, in network calculus for communication networks, the unit is typically bit and $A(t)$ denotes the total number of bits in the arrived packets up to time t [3].

5.2 Max-Plus Queueing Principle Approach

The focus of this subsection is also on deriving bounds on waiting time in queue, but based on the max-plus queueing principle. Since in $GI/GI/1$, the service time of a customer is independent of its waiting time in queue, with the bounds on queueing delay, we can further derives the corresponding bounds on system delay.

Recall the following from (33):

$$W(n) = \max_{0\le m\le n}\left[\sum_{k=m}^{n-1}\delta(k) - \sum_{k=m}^{n-1}\tau(k)\right] \quad (43)$$

where by convention, $\sum_{k=m}^{n} x(k) = 0$ if $m > n$.

Following a similar approach and based on Chernoff bound, we obtain for any $x \ge 0$ the following inequality for waiting

[3] A packet is said to have arrived when and only when its last bit has arrived.

time in queue:

$$\begin{aligned} & P\{W(n) > x\} \\ = \; & P\{\max_{0\le m\le n-1} \sum_{k=m}^{n-1} [\delta(k) - \tau(k)] > x\} \\ \le \; & e^{-\theta x} \sum_{m=0}^{n-1} E[e^{\theta \sum_{k=m}^{n-1}[\delta(k)-\tau(k)]}] \\ = \; & e^{-\theta x} \sum_{m=0}^{n-1} E[e^{\theta(\delta(0)-\tau(0))}]^{n-m} \end{aligned}$$

and then

$$P\{W > x\} \le \frac{M_{\delta(0)-\tau(0)}(\theta)}{1 - M_{\delta(0)-\tau(0)}(\theta)} e^{-\theta x} \quad (44)$$

for any $\theta > 0$, under the condition $M_{\delta(0)-\tau(0)}(\theta) < 1$.

Also similarly, a refined bound based on martingale can be derived. Particularly, fixing n, we define a stochastic process $V(l) = e^{\theta \sum_{k=n-1-l}^{n-1}[\delta(k)-\tau(k)]}$ with $\theta > 0$ and $0 \le l < n-1$. Note that $e^{\theta[\delta(n-1-(l+1))-\tau(n-1-(l+1))]}$ is independent of all $e^{\theta[\delta(n-1-v)-\tau(n-1-v)]}$ for all $v = 0, 1, \ldots, l$. Let $\mathcal{M}_l$ denote the σ-algebra generated from $\{V(l)\}$. We then have if $E[e^{\theta(\delta(0)-\tau(0))}] \le 1$:

$$\begin{aligned} E[V(l+1)|\mathcal{M}_l] &= V(l) E[e^{\theta[\delta(n-1-(l+1))-\tau(n-1-(l+1))]}] \\ &= V(l) E[e^{\theta(\delta(0)-\tau(0))}] \\ &\le V(l) \end{aligned} \quad (45)$$

This indicates that $\{V(l)\}$, $l = 0, 1, n-1$, is a supermartingale. In this way, with Doob's inequality for martingale, we get from (43):

$$\begin{aligned} & P\{W(n) > x\} \\ = \; & P\{e^{\theta \max_{0\le m\le n} \sum_{k=m}^{n-1}[s(k)-\tau(k+1)]} > e^{\theta x}\} \\ = \; & P\{\max_{0\le l<n-1} v(l) > e^{\theta x}\} \\ \le \; & E[e^{\theta v(0)}] e^{-\theta x} = E[e^{\theta(\delta(0)-\tau(0))}] e^{-\theta x}. \end{aligned} \quad (46)$$

Again, both bounds (44) and (46) have similar form as the Kingman's bound. The bound in (46) is obviously tighter than that of (44), meaning the supermartingale approach outperforms the Chernoff bound. The following theorem follows directly from (46):

THEOREM 2. *Consider a $GI/GI/1$ queue. Assume that $M_{\delta(0)-\tau(0)}(\theta)$ exists for small $\theta > 0$. Let $\theta_0 = \sup\{\theta : M_{\delta(0)-\tau(0)}(\theta) \le 1\}$. Then, we have the following bound for waiting time in queue:*

$$P\{W > x\} \le \inf_{0<\theta\le\theta_0} M_{\delta(0)-\tau(0)}(\theta) e^{-\theta x}. \quad (47)$$

5.3 Discussion

The martingale bounds are generally better than their corresponding Chernoff bounds. However, a direct comparison of the two martingale bounds seems to be difficult. In the following, we shall use $M/M/1$ as an example, for which both martingale bounds are found and compared.

Consider an $M/M/1$ queue with arrival rate parameter λ and service rate parameter μ, and $\lambda < \mu$. It is known that both $A(1)$ and $S(1)$ have Poisson distribution respectively with parameters λ and μ. Then, the right hand side of (40) becomes $e^{\lambda(e^{\theta}-1)} e^{(x+1)\mu(e^{-\theta}-1)}$ that, letting $\theta = \ln\frac{\mu}{\lambda}$, results in $P\{W > x\} \le e^{-(\mu-\lambda)x}$. In addition, for the $M/M/1$ queue, both $\tau(0)$ and $\delta(0)$ are exponentially distributed respectively with rates λ and μ. Then, the right hand side of (46) becomes $\frac{\mu}{\mu-\theta}\frac{\lambda}{\lambda+\theta} e^{-\theta x}$ where, by letting $\theta = \mu - \lambda$, we also get $P\{W > x\} \le e^{-(\mu-\lambda)x}$.

The $M/M/1$ example shows that the martingale bound obtained from the min-plus convolution queueing principle is as tight as the martingale bound from the max-plus convolution queueing principle. It is worth highlighting that $e^{-(\mu-\lambda)x}$ is also the exact distribution of $P\{D > x\}$, implying that there is only one customer service time difference between the martingale bounds and the exact solution. Indeed, for $M/M/1$, $P\{W > x\} = \rho e^{-(\mu-\lambda)x}$, comparing which with the bound above, one can see that when the utilization approaches 1, the bound approaches the exact result.

Interestingly, the right hand side of (42) may be treated as a bound on the system delay, i.e.

$$P\{D > x\} \le \inf_{0<\theta\le\theta_0} \left\{ M_{A(1)-S(1)}(\theta) \left[M_{S(1)}(-\theta)\right]^x \right\}. \quad (48)$$

This is due to that (40) holds for any time t. Since the discrete time system approximates the continuous time system by letting the time length of discretization unit approach 0, we can take $t = a(n)$ for any $C(n)$ and have $D(t) = d(n) - a(n) = D(n)$. Then, (40) becomes a bound on $D(n)$ and hence the bound on $P\{D > x\}$ is obtained. For the $M/M/1$ example above, the bound equals the exact result. Due to the similarity in deriving (40) and (46), one might conjecture that the right hand side of (47) could also be an upper bound on $P\{D > x\}$, but an insightful explanation for this is open to be found.

5.4 Related Work

In [24], an attempt was made to link network calculus with queueing theory. Particularly, the authors studied the $M/M/1$ case with a deterministic shaper enforced on the input and a deterministic service curve element enforced on the service. The study was performed through simulation and the queue length distribution was compared with the original $M/M/1$ case. Analytical study of the system was not touched. Essentially, the system studied in [24] can be considered as a special case of the systems studied in [22] where some analytical bounds can be found. However, the analysis in [22] and in a considerable part of the network calculus literature mainly relies on various arrival curve and service curve models, with little consideration or use of the underlying queueing principles.

Chernoff bound has been widely applied in (stochastic) network calculus (e.g. see [6] [1] [11] [15]). Expressing network calculus properties in the form of moment generating function was initially made in [11] where a Chernoff delay bound for the single node case is implied. In addition, the analysis in Section 5.1 resembles much the single node case analysis in [8], even though the final expression of the bounds and the condition for them have small difference in [11]. Nevertheless, there is a conceptual difference in deriving these bounds. In [8], the approach heavily relies on concepts and results of stochastic network calculus, while in Section 5.1, equipped with the min-plus convolution queueing principle, we have adopted a classical approach that has long been used in queueing theory (as early as [17]).

Due to the conceptual difference, when it comes to the application and calculation of the obtained bounds, the idea in [8] is to use a compound process to represent the arrival, allowing to represent the server using a fluid view with constant service rate. Specifically, the compound arrival process is $A(t) = \sum_{n=0}^{N(t)} \delta(n)$ where $C(N(t))$ is the last customer that has arrived by time t. For the $M/M/1$ case, $\delta(n)$ is exponentially distributed while $N(t)$ has a Poisson distribution. However, the analysis in Section 5.1 is directly on the arrival process and the service process. As is clear in the $M/M/1$ example, $A(t)$ in this paper is interpreted as the number of arrivals by time t. Later in the tandem network case, an advantage of such way of interpreting $A(t)$ will be discussed.

Using martingale approach to obtain improved bounds for queues can be traced back to [17]. The martingale delay bound in Theorem 2 is similar to the one in [27]. However, in [17] [27] and other related queueing theory literature, the obtained martingale bounds are typically based on the Lindley queueing principle, while our analysis has the root on the two network calculus queueing principles. In [5], the martingale approach has also been used to analyze linear systems under max-plus algebra, but the expression is in a matrix form.

6. TANDEM SERVERS

Having introduced the application of min-plus convolution and max-plus convolution queueing principles to $GI/GI/1$ analysis, we demonstrate in this section their application to network analysis. Particularly, we consider a network of tandem servers, each with a FIFO queue. In this system, customers first enter the first queue; after leaving the hth server, they immediately enter the $(h+1)$st server, where $h = 1, 2, \ldots, H$ and H denotes the total number of servers. We shall still adopt the notation introduced in Table 1 but with subscript h to represent the hth server, e.g. $a_h(n)$ denotes the arrival time of customer $C(n)$ to server h.

For the tandem system, we consider two general scenarios: Scenario I and Scenario II. The focus will be on the effect of changing the order of servers. We start with introducing queueing principles implied in the network.

6.1 Queueing Principles

For a tandem network, there always holds $d_h(n) = a_{h+1}(n)$, $h = 1, 2, \ldots, H-1$, by definition. Then, by iteratively applying (31) to its right hand side, we get

$$d_H(n) = ((a_1 \bar{\otimes} S_1) \bar{\otimes} \cdots) \bar{\otimes} S_H(0, n). \tag{49}$$

One may have noticed the slight difference between the right hand side of (49) and the right hand side of (14). This is due to that we have not proved the associativity of $\bar{\otimes}$ for the bivariate case. Before the proof, let us define the function family $\hat{\mathcal{F}}$:

$$\hat{\mathcal{F}} = \{\hat{\mathcal{F}}(\cdot,\cdot) : F(s,t) \geq 0, F(s,t) \leq F(s,\tau)\}, \forall 0 \leq s \leq t \leq \tau.$$

We now consider bivariate functions $F, G, H \in \hat{\mathcal{F}}$. We have the following associativity of max-plus convolution of bivariate functions: For any $0 \leq m \leq n$

$$\begin{aligned} & (F \bar{\otimes} G) \bar{\otimes} H(m,n) \\ = & \sup_{m \leq l \leq n} \sup_{m \leq k \leq l} \{F(m,k) + G(k,l) + H(l,n)\} \\ = & \sup_{m \leq k \leq n} \sup_{k \leq l \leq n} \{F(m,k) + G(k,l) + H(l,n)\} \\ = & F \bar{\otimes} (G \bar{\otimes} H)(m,n). \end{aligned} \tag{50}$$

Similarly, the following associativity can be verified for min-plus convolution of bivariate functions: For any $0 \leq s \leq t$,

$$(F \otimes G) \otimes H(s,t) = F \otimes (G \otimes H)(s,t). \tag{51}$$

With (50), we can re-write (49) as:

$$d_H(n) = a_1 \bar{\otimes} (S_1 \bar{\otimes} \cdots \bar{\otimes} S_H)(0,n). \tag{52}$$

which we call the *max-plus convolution queueing principle* for the tandem network.

Also for the tandem network, we can extend (27) and based on (51), get the *min-plus convolution queueing principle* as

$$A_H^*(t) = A_1 \otimes (S_1 \otimes \cdots \otimes S_H)(0,t). \tag{53}$$

if for all $h = 1, \ldots, H-1$, the following condition is satisfied

$$A_h^*(t) = A_{h+1}(t). \tag{54}$$

It is worth highlighting that, depending on the definition of $A(t)$, condition (54) does not necessarily hold in tandem networks as to be discussed later in Section 6.3.

6.2 Scenario I

Under Scenario I, the amount of service required by a customer $C(n)$ does not change when the customer traverses the network. Let us denote by $l(n)$ such service amount. The service rate of each server h, in terms of the amount of service the server can provide per unit time, is assumed to be constant and denoted by r_h. Accordingly, we have $\delta_h(n) = \frac{l(n)}{r_h}$, i.e. the service time of customer $C(n)$ at server h. Scenario I is typical in modern packet-switched communication networks, where $C(n)$ represents a packet level customer, i.e. a packet, and $l(n)$ is simply the length (in bits) of the packet, which does not change when the packet traverses the network.

PITFALL 1. *Scenario I can sometimes lead to a pitfall. According to the setting, each server has a constant service rate r_h. One might then think each server would provide a service curve $\beta_h(t) = r_h \cdot t$, and the network provide a network service curve $\beta_H(t) = (\min_h r_h)t$ according to (11). In the discrete time case, one could instead use $\beta_h(t) = r_h \cdot (t-1)^+$ and $\beta_H(t) = (\min_h r_h)(t-H+1)^+$ by considering discretization effect and ignoring the packetizer at the last server. One could further conclude results from (11) and the single node analysis by treating the whole network as a single server with service curve β_H. One such conclusion could be that changing the order of servers in the network would not affect the end-to-end performance owing to the commutativity of min-plus convolution.*

To show the above is indeed a pitfall, let us consider a simple example. In this example, we consider a network of two servers. Suppose the first server has service rate $r_1 = 3$ and the second server has service rate $r_2 = 2$. In addition,

there are only two customers arriving to the network back-to-back. For the first customer $C(0)$, $l(0) = 4$; and for the second customer, $l(1) = 2$. Without loss of generality, we assume the arrival times of these two customers are $a(0) = a(1) = 0$. Under this setting, one can verify that the two customers leave the network respectively at $d_2(0) = \frac{10}{3}$ and $d_2(1) = \frac{13}{3}$. However, according to the discussion above, the network would have a service curve $\beta_H(t) = 2(t-1)^+$ and hence a bound 4 on maximum delay of any packet, which is smaller than $\frac{13}{3}$ and hence unfortunately wrong. Now, let us change the order of the two servers. In the resulting network, the first server has service rate 2 and the second server has service rate 3. Similarly we consider two customer arrivals with the same parameters. Under this setting, it can be verified that $d'_2(0) = \frac{10}{3}$ and $d'_2(1) = 4$, where the bound 4 coincidentally holds.

The above example implies that either $\beta_h(t) = r_h \cdot t$ or $\beta_h(t) = r_h \cdot (t-1)^+$ is not a service curve. A correct service curve of a constant rate server is not only determined by the rate parameter but also dependent on the (maximum) required service by a customer, e.g. packet length when a packet-switched network is concerned. Note that the pitfall also applies when there are crossing customers at servers in the network.

In addition, changing the order of servers in the tandem network can cause the end-to-end delay changed as well. The underlying reason is implied in (53) where the min-plus convolution of bivariate functions is **not commutative** as oppose to the commutativity of min-plus convolution of single variate functions e.g. in (11).

Nevertheless, are there cases where changing the order of servers does not change the end-to-end delay? The answer is positive and in the following, we show two such cases.

One case is that the service time of a customer remains the same at all servers, i.e., $\delta_1(k) = \cdots = \delta_H(k) = \delta(k)$ for any customer $C(k)$ even though $\delta(k)$ may itself be random. An example is that the network is packet-switched. The packet interarrival times are i.i.d., following some general distribution. The packet lengths are also i.i.d. and follow some general distribution. At the first node, it is $GI/GI/1$. All nodes have the same constant service rate, i.e. $r_1 = \cdots = r_H$. For this case, it is trivial that changing the order of servers does not change the end-to-end delay, since these servers behave the same and may be numbered arbitrarily. Formal proof can also be formulated e.g. by applying Lemma 1 and Lemma 2.

Another case is that all customers have the same required service at any server, each server is a constant rate server, and the service rates of these servers may be different. Again we can use the packet-switched network as an example. At the input, the packet interarrival times are i.i.d., following some general distribution. However, all packets have the same length l. In addition, all servers are constant rate servers but their service rates r_h, $h = 1, \cdots, H$, may be different. For this case, the following exciting result holds. Its proof is included in Appendix.

THEOREM 3. *Consider a network of tandem FIFO servers. Customers enter the network from the first server. At the input, customer interarrival times are i.i.d. and all customers have the same amount of required service. Each server is a constant rate server. Then, changing the order of servers in the network does not affect the end-to-end delay and network backlog performance.*

6.3 Scenario II

Under Scenario II, the service times of customers at each server are i.i.d., and are independent of customer service times at other servers. For this scenario, it is worth highlighting that while the max-plus queueing principle extends naturally to the network case, special care is needed when one tries to apply the min-plus queueing principle:

Recall that in network calculus, $A_h(t)$ typically denotes the traffic amount (in bits) of packets that have arrived to node h by time t. Suppose the nth packet is the latest packet arrival to node h by time t. Then, we can write $A_h(t)$ as follows:

$$A_h(t) = \sum_{m=1}^{n} l(m) \tag{55}$$

where $l(m)$ denotes the length of the mth packet.

PITFALL 2. *For Scenario II, if one were to adopt (55) as the way of interpreting $A_h(t)$, i.e. $A_h(t) = \sum_{m=1}^{n} l_h(m)$ where $l_h(m)$ denotes the required service amount by customer m at server h, and still rely the analysis on the tandem network min-plus convolution queueing principle, there would be a problem. This is due to that we now do not have $l_{h-1}(m) = l_h(m)$ nor $A^*_{h-1}(t) = A_h(t)$, and consequently will lose the promising concatenation property.*

A simple fix to the above problem is to interpret $A_h(t)$ as the *number* of arrivals up to time t. Under this new way of interpreting, $A^*_{h-1}(t) = A_h(t)$ holds for Scenario II, and so does the concatenation property.

Scenario II has been widely studied in the context of queueing theory. For example, when the arrivals follow a Poisson distribution and the service times at each server are exponentially distributed, the tandem network becomes an open Jackson network. For such a network, the queueing theory literature tells that the network acts as if each node could be viewed as an independent $M/M/1$ queue. An immediate implication is that changing the order of servers in the tandem does not affect the end-to-end network performance.

Figure 1 presents simulation results of a network of 3 nodes in tandem. Particularly, it displays the complementary cumulative distribution function (CCDF) of end-to-end delay under different arrival and service patterns. The following arrival/service time combinations are studied: (a) *M/M/1*: Poisson arrival at the input, and exponentially distributed service time at each node; (b)*M/Uniform/1*: Poisson arrival at the input and the service time has uniform distribution respectively with the range in [0.1, 1.9], [0.2, 2], [0.3, 2.1] for the nodes; (c) *Erlang/Uniform/1*: At the input, the customer interarrival time follows Erlang-3 distribution and the service time is as in (b); (d) *Hyper-exponential /Uniform /1*: At the input, the customer interarrival time follows hyper-exponential distribution with order 2 and branch probability {0.3, 0.7} and the mean on each branch {2.5, 0.84}, and the service time distribution is as in (b).

In addition, we consider two cases where the order of nodes in the network is changed. One is denoted by "incre" on the figure, where the load of the 3 servers is $(0.75, 0.825, 0.9)$. Another is denoted by "decre" in which case, the order of the 3 nodes is reversed and correspondingly the load of the 3 servers becomes $(0.9, 0.825, 0.75)$.

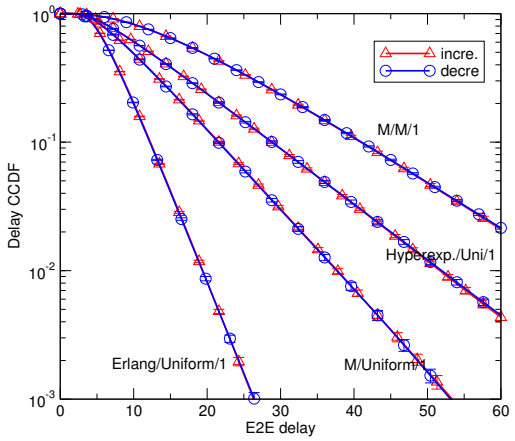

Figure 1: Tandem network

Figure 1 shows that there is an exact match between "incre" and "decre" under the *M/M/1* combination as expected. Interestingly, under all other combinations, such a match is also found. This motivated us to study if this phenomenon exists for the general setting. The answer is positive and the conclusion is given in the following theorem.

THEOREM 4. *Consider a network of tandem FIFO servers. Customers enter the network from the first server. Customer interarrival times are i.i.d. At each server, the service times are also i.i.d. and are independent of services times at all other servers. Then, changing the order of servers in the network does not affect the end-to-end delay and network backlog performance.*

The following result is critical in proving Theorem 4. The detailed proof of Theorem 4 is in included in Appendix.

LEMMA 3. *Consider two consecutive nodes in the tandem. There holds:*

$$\Delta_h \bar{\otimes} \Delta_{h+1}(m,n) \quad =_{st} \quad \Delta_{h+1} \bar{\otimes} \Delta_h(m,n). \qquad (56)$$

6.4 Discussion

The implication of the investigation in this section is twofold. One is that much care is needed when applying network calculus and some existing results need re-examination. Particularly, Pitfall 1 has found easy success in both deterministic and stochastic network calculus. In stochastic network calculus, the situation seems to be even worse. For example, in Ch. 6.2.6 of [15] co-authored by this author and in some representative works of stochastic network calculus, the (stochastic) service curve of a constant rate link is often simply represented as $R \cdot (t-s)$ where R denotes the rate of the link. If there is cross traffic, a widely used leftover service curve in them is $[R \cdot (t-s) - A_c(s,t)]^+$, where $A_c(s,t)$ denotes the amount of cross traffic in $(s,t]$. In both cases, the packetization effect on the (stochastic) service curve to the considered input is not well taken into account. It is worth highlighting that this does not affect single node analysis results due to a property of packetizer, i.e. the packetization effect can be ignored for the last node on the path of the considered input [6] [19]. However, for network case analysis, the packetization effect on the considered input must be taken into account in choosing proper stochastic service curves.

Another implication is that Theorems 3 and 4 can be made use of to analyze end-to-end network performance of the corresponding network. The following results follow from Theorems 3. Similar analysis can be made based on Theorem 4 and the idea in its proof, which is left as our future work.

COROLLARY 1. *Consider the same tandem network as Theorem 3. The end-to-end queueing delay and backlog in queue performance of the network are the same as of a single server queue that has the (stochastically) same input and has the service rate equal to the lowest server service rate in the tandem network.*

For the considered network in Theorem 3, it can be verified that after arranging servers in the increasing service rate order in the tandem, there is no waiting time or waiting customer in queue at nodes $h' = 2, \ldots, H$. In other words, in the ordered tandem, queueing delay and backlog are only seen at the first node $h' = 1$, and hence Corollary 1 follows.

Further with Corollary 1 and the single server queue analysis in Section 5, we conclude the following delay bound for the tandem network, which is based on the max-plus martingale bound. Similar bound can also be obtained from the min-plus martingale bound in Section 5.

COROLLARY 2. *Consider the same tandem network as Theorem 3. Let $\delta = \frac{l}{\min_h r_h}$, where l denotes the required service amount of each customer and r_h the average service rate of server h, $h = 1, \ldots, H$. Assume that $M_{\delta-\tau(0)}(\theta)$ exists for small $\theta > 0$ and let $\theta_0 = \sup\{\theta : M_{\delta-\tau(0)}(\theta) \leq 1\}$. Then, the network queueing delay W_{net} satisfies:*

$$P\{W_{net} > x\} \leq \inf_{0<\theta\leq\theta_0} M_{\delta-\tau(0)}(\theta) e^{-\theta x}. \qquad (57)$$

7. SUMMARY

We introduced two fundamental queueing principles that underlie network calculus, which are the min-plus convolution queueing principle and the max-plus convolution queueing principle. They compliment the well-known Lindley recursion queueing principle. Based on the two network calculus queueing principles, we derived delay bounds for the single node case, which are consistent with similar bounds derived based on the Lindley recursion queueing principle. In addition, we extended the network calculus queueing principles to the tandem network case, and proved that under some general conditions, changing the server order in a tandem does not affect the end-to-end network performance. This result is fundamental and may be further explored to derive performance bounds, e.g. delay bound and backlog bound, using similar approach as used for the single node case. It should be stressed that network calculus, particularly stochastic network calculus, has results that can be readily explored to find inequalities for classical queueing problems. However, for this paper, we have intentionally chosen to focus on deriving results directly from the network calculus queueing principles. Essentially, we demonstrated that the network calculus queueing principles allow to study classical queueing problems and derive results, which might otherwise be difficult to obtain from only using the Lindley recursion queueing principle, from a different perspective. This not only provides new insights on queue analysis but also link network calculus to queueing theory.

Acknowledgment

The author would like to thank Guoqiang Hu for discussion and producing the figure and thank the anonymous reviewer for helpful comments.

APPENDIX

A. PROOF OF THEOREM 3

We shall only prove the delay part. The backlog part follows similarly.

The essential idea is to compare the end-to-end delay performance of two tandem networks. In one network denoted as NET1, servers are ordered in the sequence as $S_1, S_2, \ldots, S_H$. In the other network denoted as NET2, the order of servers is changed and we mark the new order as $S'_1, S'_2, \ldots, S'_H$ where the set of servers $\{S'_1, \ldots, S'_H\}$ are the same as $\{S_1, \ldots, S_H\}$. All servers are constant rate server. We denote by r_h and r'_h the service rate of S_h and S'_h respectively. Denote by $a_h(n)$ the arrival time of customer $C(n)$ to the hth server in the first network, and by $a'_h(n)$ the arrival time of customer $C'(n)$ to the hth server in the second network . All customers have the same amount of required service in both networks, denoted by L. Correspondingly, the customer service times at any server are constant, and we shall denote by δ_h and δ'_h the customer service time at S_h and S'_h respectively. The inputs to both networks have the same i.i.d. interarrival times denoted by $\{\tau(0), \tau(1), \ldots\}$ and $\{\tau'(0), \tau'(1), \ldots\}$. By assumption, we have $\tau(n) =_{st} \tau'(n)$ for all $n = 0, 1, 2, \ldots$.

Consider NET1. The following commutativity holds:

$$\begin{aligned} & \Delta_h \bar{\otimes} \Delta_{h+1}(m,n) \\ = & \sup_{m \le l \le n} [(l-m+1)\delta_h + (n-l+1)\delta_{h+1}] \\ = & \sup_{m \le k \le n} [(k-m+1)\delta_{h+1} + (n-k+1)\delta_h] \\ = & \Delta_{h+1} \bar{\otimes} \Delta_h(m,n) \end{aligned} \tag{58}$$

which implies that we can arbitrarily change the order S_h, $h = 1, \ldots, H$, in the max-plus convolution principle for tandem networks. Let us change the order of servers in NET1 such that the resulting order is the same as the corresponding servers in NET2 and get:

$$d_H(n) = a_1 \bar{\otimes} (\Delta'_1 \bar{\otimes} \cdots \bar{\otimes} \Delta'_H)(0,n), \tag{59}$$

Let us define the following delay function

$$\begin{aligned} & D_n(X_1, \ldots, X_n) \\ = & \sup_{0 \le m \le n} [\sum_{k=m}^{n-1} X_k + \Delta'_1 \bar{\otimes} \cdots \bar{\otimes} \Delta'_H(m,n)]. \end{aligned} \tag{60}$$

It can be easily verified that the function is nondecreasing in $\{x_1, \ldots, x_n\}$.

Then, for end-to-end delay in NET1, we have

$$\begin{aligned} D(n) &= d_H(n) - a_1(n) \\ &= \sup_{0 \le m \le n} [a_1(m) - a_1(n) + \Delta'_1 \bar{\otimes} \cdots \bar{\otimes} \Delta'_H(m,n)] \\ &= \sup_{0 \le m \le n} [-\sum_{k=m}^{n-1} \tau(k) + \Delta'_1 \bar{\otimes} \cdots \bar{\otimes} \Delta'_H(m,n)]. \\ &= D_n(-\tau(0), \ldots, -\tau(n-1)) \end{aligned} \tag{61}$$

For delay in NET2, we have:

$$D'(n) = D_n(-\tau'(0), \ldots, -\tau'(n-1)). \tag{62}$$

Since both $\tau(k)$ and $\tau'(k)$, $k = 0, 1, \ldots, n$, are i.i.d. random variables and $-\tau_k =_{st} -\tau'_k$, we then have from Lemma 1:

$$\{-\tau'(0), \ldots, -\tau'(n-1)\} =_{st} \{-\tau(0), \ldots, -\tau(n-1)\}.$$

Further from Lemma 2, we can conclude:

$$D(n) =_{st} D'(n)$$

which ends the proof.

B. PROOF OF THEOREM 4

We only prove the delay part, since the backlog part follows similarly.

For the proof, we adopt the same idea as used above, which is to compare the end-to-end delay performance of two tandem networks. In one network denoted as NET1, servers are ordered in the sequence as $S_1, S_2, \ldots, S_H$. In the other network denoted as NET2, the order of servers is changed and we mark the new order as $S'_1, S'_2, \ldots, S'_H$ where each of these S'_h in NET2 has a counterpart $S_{h'}$ in NET1. S'_h and $S_{h'}$ have the same service characteristics in the sense that both of them have i.i.d. service times $\delta'_h(n)$ and $\delta_{h'}(n)$, $n = 0, 1, 2, \ldots$. In addition, $\delta'_h(n) =_{st} \delta_{h'}(n)$, for all $n = 0, 1, 2, \ldots$. Furthermore, the inputs to both networks have the same i.i.d. interarrival times denoted by $\{\tau(0), \tau(1), \ldots\}$ and $\{\tau'(0), \tau'(1), \ldots\}$. By assumption, we have $\tau(n) =_{st} \tau'(n)$ for all $n = 0, 1, 2, \ldots$.

Consider NET1. We have the following commutativity in the stochastic equality sense:

$$\begin{aligned} & \Delta_h \bar{\otimes} \Delta_{h+1}(m,n) \\ = & \sup_{m \le l \le n} [\sum_{k=m}^{l} \delta_h(k) + \sum_{k=l}^{n} \delta_{h+1}(k)] \\ =_{st} & \sup_{m \le l \le n} [\sum_{k=n+m-l}^{n} \delta_h(k) + \sum_{k=m}^{n+m-l} \delta_{h+1}(k)] \quad (63) \\ = & \sup_{m \le l' \le n} [\sum_{k=m}^{l'} \delta_{h+1}(k) + \sum_{k=l'}^{n} \delta_h(k)] \quad (64) \\ = & \Delta_{h+1} \bar{\otimes} \Delta_h(m,n), \quad (65) \end{aligned}$$

where step (64) is obtained by letting $l' = n+m-l$. Step (63) is critical. To see how it is obtained, let us define function $F(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \sup_{m \le l \le n} [\sum_{k=m}^{l} X_k + \sum_{k=l}^{n} Y_k]$ that is clearly nondecreasing in $\{x_1, \ldots, x_n, y_1, \ldots, y_n\}$. Letting $X_i = \delta_h(i)$ and $Y_i = \delta_{h+1}(i)$, $i = 1, \ldots, n$ results in the form before $=_{st}$. Letting $X_i = \delta_h(n-i)$ and $Y_i = \delta_{h+1}(n-i)$ instead, we get the right hand side on $=_{st}$ at step (63). Since $\delta_h(i)$, $i = 0, 1, \ldots, n$, are i.i.d. and so are $\delta_{h+1}(i)$, the stochastic equality follows from Lemma 1 and Lemma 2.

Since all servers are independent of each other, (65) implies that if we could change the order S_h, $h = 1, \ldots, H$ in the max-plus convolution principle for NET1, the resulting network, which has the same input as NET1, would have the same performance stochastically as the original NET1. Following this, let us change the order of servers in NET1 such that the resulting order is the same as the corresponding servers in NET2 and get:

$$d''_H(n) = a_1 \bar{\otimes} (\Delta''_1 \bar{\otimes} \cdots \bar{\otimes} \Delta''_H)(0,n). \tag{66}$$

The stochastic commutativity (65) implies that $d_H(n) =_{st} d''_H(n)$ and hence the end-to-end delay in NET1 satisfies

$$D(n) =_{st} d''_H(n) - a_1(n) \equiv D''(n). \tag{67}$$

Let us now define the following delay function

$$\begin{aligned} & D_n(X_1,\ldots,X_n,Y_{1,1},\ldots,Y_{1,n},\ldots,Y_{n,1},\ldots,Y_{n,n}) \\ = & \sup_{0\le n_H\le n}\cdots\sup_{0\le n_1\le n_2}\{\sum_{k_1=n_1}^{n_2-1}[Y_{1,k_1}+X_{k_1+1}]+ \\ & \cdots+\sum_{k_H=n_H}^{n-1}[Y_{H,k_H}+X_{k_H+1}] \\ & +[Y_{1,n_2}+\cdots+Y_{H,n}]\} \end{aligned} \quad (68)$$

which is nondecreasing in

$$\{x_1,\ldots,x_n,y_{1,0},\ldots,y_{1,n},\ldots,y_{H,0},\ldots,y_{H,n}\}.$$

Then, for end-to-end delay in NET1, we have from (67)

$$\begin{aligned} & D''(n) = \\ & D_n(-\tau(0),\ldots,-\tau(n-1),\ldots,\delta_h''(0),\ldots,\delta_h''(n),\ldots) \end{aligned} \quad (69)$$

For delay in NET2, we have:

$$\begin{aligned} & D'(n) = \\ & D_n(-\tau'(0),\ldots,-\tau'(n-1),\ldots,\delta_h'(0),\ldots,\delta_h'(n),\ldots) \end{aligned} \quad (70)$$

Note that all $\tau(k)$ and $\tau'(k)$, $\delta_h''(k)$ and $\delta_h'(k)$ $(h=1,\ldots,H)$, $k=1,\ldots,n$, are i.i.d. random variables and $\tau_k =_{st} \tau_k'$, $\delta_h''(k) =_{st} \delta_h'(k)$ for all $h=1,\ldots,H$ and $k=0,1,\ldots,n$, we can then conclude from Lemma 1 and Lemma 2:

$$D'(n) \quad =_{st} \quad D''(n) =_{st} D(n) \quad (71)$$

which ends the proof.

C. REFERENCES

[1] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Statistical service assurances for traffic scheduling algorithms. *IEEE J. Select. Areas Commun.*, 18(12):2651–2664, 2000.

[2] E. Brockmeyer, H. L. Halstrøm, and A. Jensen. The life and works of A.K. Erlang. *Transactions of the Danish Academy of Technical Sciences*, 2, 1948.

[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. On the exponentiality of stochastic linear systems under the max-plus algebra. *IEEE Trans. Automatic Control*, 41(8):1182–1188, August 1996.

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

[7] C.-S. Chang and Y. H. Lin. A general framework for deterministic service guarantees in telecommunication networks with variable length packets. *IEEE Trans. Automatic Control*, 46:210–221, 2001.

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

[9] 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.

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

[12] P. Goyal, S. S. Lam, and H. M. Vin. Determining end-to-end delay bounds in heterogeneous networks. In *Proc. NOSSDAV* 1995.

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

[14] Y. Jiang. Network calculus - bridging erlang and the internet. In *Annual Report 2008, Q2S Center, NTNU*, 2009.

[15] Y. Jiang and Y. Liu. *Stochastic Network Calculus*. Springer, 2008.

[16] 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.

[17] J. Kingman. A martingale inequality in the theory of queues. *Proc. Camb. Phil. Soc.*, 59:359–361, 1964.

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

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

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

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

[22] Y. Liu, C.-K. Tham, and Y. Jiang. Conformance analysis in networks with service level agreements. *Computer Networks*, 47:885–906, 2005.

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

[24] K. Pandit, J. Schmitt, and R. Steinmetz. Network calculus meets queueing theory - a simulation based approach to bounded queues. In *Proc. IWQoS* 2004.

[25] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in intergrated services networks: The single-node case. *IEEE/ACM Trans. Networking*, 1(3):344–357, 1993.

[26] D. Stoyan. *Comparison Methods for Queues and Other Stochastic Models*. John Wiley & Sons, 1983.

[27] H. H. Tan. Another martingale bound on the waiting-time distribution in GI/G/1 queues. *Journal of Applied Probability*, 16(2):454–457, 1979.

[28] J. Xie and Y. Jiang. Stochastic service guarantee analysis based on time-domain models. In *Proc. MASCOTS* 2009.

[29] O. Yaron and M. Sidi. Performance and stability of communication network via robust exponential bounds. *IEEE/ACM ToN*, 1:372–385, 1993.

[30] L. Zhang. Virtual clock: a new traffic control algorithm for packet switching networks. In *Proc. ACM SIGCOMM* 1990.