A General Framework for Deterministic Service Guarantees in Telecommunication Networks with Variable Length Packets

Cheng-Shang Chang* Yih Haur Lin†

(May 1998; Revised: September 1999; Final: June 2000)

Abstract

By extending the recently developed filtering theory under the min-plus algebra to the max-plus algebra, in this paper we develop a general framework for providing deterministic quality-of-service guarantees in telecommunication networks with variable length packets. The traffic in such networks is modelled by marked point processes that consist of two sequences of variables: the arrival times and the packet lengths. By modifying Cruz's characterization for a counting process, we propose a traffic characterization, called g-regularity, to characterize a marked point process. Based on the traffic characterization, we introduce two basic network elements: (i) traffic regulators that generate g-regular marked point processes, and (ii) g-servers that provides quality-of-service for marked point processes. As in the filtering theory under the min-plus algebra, network elements can be joined by concatenation, "filter bank summation," and feedback to form a composite network element. We illustrate the use of the framework by various examples that include $G/G/1$ queues, VirtualClock, guaranteed rate servers in tandem, segmentation and reassembly, jitter control, dampers, and window flow control.
Keywords: marked point processes, service guarantees, max-plus algebra, traffic regulation

*Corresponding author. Dept. of Electrical Engineering, National Tsing Hua University, Hsinchu 300, Taiwan, R.O.C. Email: cschang@ee.nthu.edu.tw. This research is supported in part by the National Science Council, Taiwan, R.O.C., under Contract NSC-87-2213-E007-084. The conference version that contains part of the results in this paper was presented at IEEE IWQoS'98.

†Dept. of Electrical Engineering, National Tsing Hua University, Hsinchu 300, Taiwan, R.O.C. Email: yh-lin@gibbs.ee.nthu.edu.tw

1 Introduction

In [9, 10], Cruz first demonstrated that deterministic service guarantees can be achieved in a telecommunication network provided that the traffic entering a network is well regulated. He proposed using the following traffic characterization: a counting process A (with $A(t)$ being the cumulative number of arrivals by time t) is f-upper constrained for some function f if

$$A(t) - A(s) \leq f(t-s), \quad s \leq t. \tag{1}$$

Based on such a traffic characterization, a filtering theory is recently developed for deterministic traffic regulation and service guarantees (see e.g., [5, 1, 14, 21]). The filtering theory is developed under the $(\min, +)$-algebra, where one replaces the usual addition by the min operator and the usual multiplication by the addition operator. As in the classical linear system theory, the new filtering theory treats an arrival process A (or a departure process B) as a signal, and a network as a system. There are two types of basic network elements: the maximal f-regulator and the f-server, where the maximal f-regulator performs traffic regulation and the f-server guarantees quality-of-service. Both network elements can be viewed as linear time invariant filters with the impulse response f under the $(\min, +)$-algebra. As linear time invariant filters in the classical filtering theory, network elements can be joined by concatenation, "filter bank summation", and feedback to form a composite network element. The whole network, viewed as a composite network element, can then be analyzed by finding its impulse response under the $(\min, +)$-algebra.

The filtering theory under the $(\min, +)$-algebra is most suitable for continuous fluid networks or telecommunication networks with fixed packet size, e.g., the ATM networks. For networks with variable length packets, e.g., the Internet, additional efforts must be taken to deal with the problem of packetization. Instead of using packetizers as in [19, 13], in this paper we propose a direct treatment for providing deterministic quality-of-service guarantees in telecommunication networks with variable length packets. In order to do so, we use marked point processes to model the traffic in such networks. A marked point process $\Psi = (\tau, \ell)$ consists of two sequences of variables $\tau = \{\tau(n), n = 0, 1, 2 \ldots\}$ and $\ell = \{\ell(n), n = 0, 1, 2, \ldots\}$, where $\tau(n)$ and $\ell(n)$ are the arrival time and the packet length of the $n + 1^{th}$ packet. To characterize a marked point process, we modify Cruz's traffic characterzation as follows: a marked point process $\Psi = (\tau, \ell)$ is g-regular if

$$\tau(n) - \tau(m) \geq g(L(n) - L(m)), \tag{2}$$

where $L(n) = \sum_{m=0}^{n-1} \ell(m)$ is the sum of the packet lengths of the first n arrivals and g is an increasing sequence with $g(0) = 0$. Note that the traffic characterization in (2) provides a lower bound on the distance between two arrivals, while the traffic characterization in (1) provides an upper bound on the number of arrivals within an interval. These two traffic characterzations are related by the inverse mapping as we shall see in Section 3.4.

Based on the traffic characterization in (2), we extend the filtering theory under the $(\min, +)$-algebra to the $(\max, +)$-algebra by introducing two basic network elements: (i) traffic regulators that generate g-regular marked point processes, and (ii) g-servers that provide quality-of-service for marked point processes. As in the filtering theory, we develop the rules for network elements to be joined by concatenation, "filter bank summation," and feedback to form a composite network element. We illustrate the use of these rules by various examples that include $G/G/1$ queues, guaranteed rate servers in tandem [20], segmentation and reassembly, jitter control, dampers [12], and window flow control [1, 14, 5]. Our generalization of the filtering theory from the $(\min, +)$-algebra to the $(\max, +)$-algebra not only provides a direct, simple, and general treatment for service guarantees in networks with variable length packets, but also establishes the connections between the filtering theory and many service guarantee frameworks in the literature, e.g., VirtualClock [27, 17, 26], and guaranteed rate scheduling [20].

The paper is organized as follows. In Section 2, we briefly introduce some properties of the $(\max, +)$-algebra that will be used in the paper. In Section 3, we discuss traffic characterization and regulation for marked point processes. In Section 4, we discuss service guarantees for marked point processes. This includes the introduction of g-servers and its associated calculus. We conclude the paper in Section 5 by discussing possible extensions of the framework.

2 Preliminaries on the max-plus algebra

In this section, we briefly introduce some properties of the max-plus algebra that will be used in this paper. Let $\mathcal{F}_{inc}$ be the set of increasing (extended) real-valued sequences. That is, a sequence $f \equiv \{f(n), n = 0, 1, 2, \ldots\} \in \mathcal{F}_{inc}$ satisfies $f(m) \leq f(n)$ for all $m \leq n$. Also let $\mathcal{F}_{inc}^0 = \{f : f \in \mathcal{F}_{inc}, f(0) = 0\}$ be the set of increasing real-valued sequences with $f(0) = 0$. Note that if $f \in \mathcal{F}_{inc}^0$, then f is nonnegative as $f(n) \geq f(0) = 0$. For any two sequences f and g in $\mathcal{F}_{inc}$, we say $f = g$ (resp. $f \leq g$) if $f(n) = g(n)$ (resp. $f(n) \leq g(n)$) for all $n \geq 0$.

Consider the following two operations for sequences under the $(\max, +)$-algebra.

(i) (max) the pointwise maximum of two sequences f and g:

$$(f \oplus g)(n) = \max[f(n), g(n)].$$

(ii) (convolution) the convolution of two sequences f and g under the $(\max, +)$-algebra with respect to an *integer-valued* sequence $L \in \mathcal{F}_{inc}$:

$$(f \star_L g)(n) = \max_{0 \le m \le n} [f(m) + g(L(n) - L(m))],$$

with the convention that $-\infty + \infty = -\infty$. In particular, if $L(n) = n$, we simply write $\star_L$ as $\star$.

Clearly, these two operations are closed in $\mathcal{F}_{inc}$. Moreover, we have the following properties:

1. (Associativity) $\forall f, g, h \in \mathcal{F}_{inc}$,

$$(f \oplus g) \oplus h = f \oplus (g \oplus h).$$

2. (Commutativity) $\forall f, g \in \mathcal{F}_{inc}$,

$$f \oplus g = g \oplus f.$$

3. (Idempotency of addition) $\forall f \in \mathcal{F}_{inc}$,

$$f \oplus f = f.$$

4. (Distributivity) $\forall f, g, h \in \mathcal{F}_{inc}$,

$$(f \oplus g) \star_L h = (f \star_L h) \oplus (g \star_L h),$$
$$h \star_L (f \oplus g) = (h \star_L f) \oplus (h \star_L g).$$

5. (Monotonicity)

$$f \le f \oplus g \le \tilde{f} \oplus \tilde{g},$$
$$f \star_L g \le \tilde{f} \star_L \tilde{g},$$

for $f \le \tilde{f}$ and $g \le \tilde{g}$.

Lemma 2.1 *For $f, g, h \in \mathcal{F}_{inc}$,*

$$(f \star_L g) \star_L h \leq f \star_L (g \star h).$$

with equality holding if

$$L(n+1) - L(n) \leq 1 \tag{3}$$

for all n.

As $L(n)$ is integer-valued, the condition in (3) is equivalent to $L(n+1) = L(n)$ or $L(n+1) = L(n) + 1$ for all n.

Proof Note that

$$((f \star_L g) \star_L h)(n) = \max_{0 \leq m \leq n} \Big[\max_{0 \leq u \leq m} [f(u) + g(L(m) - L(u))] + h(L(n) - L(m)) \Big]$$
$$= \max_{0 \leq u \leq n} \Big[f(u) + \max_{u \leq m \leq n} [g(L(m) - L(u)) + h(L(n) - L(m))] \Big]. \tag{4}$$

Using the change of variable $v = L(m) - L(u)$ yields

$$\max_{u \leq m \leq n} [g(L(m) - L(u)) + h(L(n) - L(m))] \leq \max_{0 \leq v \leq L(n) - L(u)} [g(v) + h(L(n) - L(u) - v)] \tag{5}$$

with equality holding if $L(m+1) = L(m)$ or $L(m+1) - L(m) = 1$ for all $u \leq m \leq n$. Since the right hand side of (5) is $(g \star h)(L(n) - L(u))$, we have from (4) that

$$\Big((f \star_L g) \star_L h\Big)(n) \leq \Big(f \star_L (g \star h)\Big)(n)$$

with equality holding if $L(m+1) - L(m) \leq 1$ for all m. ∎

In general, the operator $\star_L$ is not associative. However, Since $L(n+1) - L(n) = 1$ for the operator $\star$, the operator $\star$ is associative, i.e.,

$$(f \star g) \star h = f \star (g \star h).$$

The operator $\star$ is also commutative, i.e.,

$$f \star g = g \star f.$$

Let $\mathbf{I}_w$ be the sequence with $\mathbf{I}_w(n) = -\infty$, $n < w$ and $\mathbf{I}_w(n) = 0$, $n \geq w$. Note that $\mathbf{I}_0(n) = 0$ for all n and that $\mathbf{I}_\infty(n) = -\infty$ for all n. Moreover, for any $g \in \mathcal{F}_{inc}$

$$\mathbf{I}_\infty \oplus g = g,$$
$$\mathbf{I}_\infty \star g = \mathbf{I}_\infty,$$
$$\mathbf{I}_0 \star g = g.$$

Thus, the sequences $\mathbf{I}_0$ and $\mathbf{I}_\infty$ are the (multiplication) identity element and the absorbing (addition) zero element in the dioid $(\mathcal{F}_{inc}, \oplus, \star)$ (see [2]). Also, we note that

$$\mathbf{I}_{w_1} \star \mathbf{I}_{w_2} = \mathbf{I}_{w_1+w_2} \tag{6}$$

and that

$$(g \star \mathbf{I}_w)(n) = \begin{cases} g(n-w) & \text{if } n \geq w \\ -\infty & \text{otherwise} \end{cases}. \tag{7}$$

Intuitively, one may view $\mathbf{I}_w$ as a shift operator for a sequence g.

A sequence $g \in \mathcal{F}_{inc}$ is superadditive if for all $m \leq n$

$$g(m) + g(n-m) \leq g(n).$$

Clearly, for a sequence $g \in \mathcal{F}^0_{inc}$, it is superadditive if and only if $g \star g = g$. If both $g_1 \in \mathcal{F}^0_{inc}$ and $g_2 \in \mathcal{F}^0_{inc}$ are superadditive, then $g_1 \star g_2$ is also superadditive as

$$(g_1 \star g_2) \star (g_1 \star g_2) = (g_1 \star g_1) \star (g_2 \star g_2) = g_1 \star g_2.$$

For any sequence $g \in \mathcal{F}_{inc}$, define the m^{th} self convolution of g by the following recursive equations:

$$\begin{aligned} g^{(1)} &= g, \\ g^{(m)} &= g^{(m-1)} \star g, \quad m \geq 2. \end{aligned}$$

As in [5], we define the *closure* of g, denoted by g^*, as follows:

$$g^* = \lim_{m \to \infty} (\mathbf{I}_0 \oplus g)^{(m)}.$$

Since the operator $\oplus$ is idempotent,

$$g^* = \lim_{m \to \infty} (\mathbf{I}_0 \oplus g \oplus g^{(2)} \oplus \ldots \oplus g^{(m)}). \tag{8}$$

In particular, for any $w > 0$ and $g \in \mathcal{F}_{inc}$, the closure of $g \star \mathbf{I}_w$ can be represented in finite terms as follows:

$$(g \star \mathbf{I}_w)^*(n) = \Big(\mathbf{I}_0 \oplus (g \star \mathbf{I}_w) \oplus \ldots \oplus (g^{(m)} \star \mathbf{I}_{mw})\Big)(n), \quad mw \leq n < (m+1)w. \tag{9}$$

This is due to the fact that

$$(g \star \mathbf{I}_w)^{(m+1)}(n) = (g^{(m+1)} \star \mathbf{I}_{(m+1)w})(n) = -\infty$$

for $n < (m+1)w$ (cf. (6) and (7)).

For more properties of the max-plus algebra, we refer to [2].

3 Traffic characterization and regulation for marked point processes

3.1 Traffic characterization

An arrival process to a telecommunication network is usually characterized by two sequences of variables: the arrival times and the service requirements (packet lengths). This can be represented by a marked point process $\Psi = (\tau, \ell)$ with $\tau = \{\tau(n), n = 0, 1, 2, \ldots\}$ and $\ell = \{\ell(n), n = 0, 1, 2, \ldots\}$, where $\tau(n)$ is the arrival time of the $n+1^{th}$ customer (packet) and $\ell(n)$ is its service requirement (packet length). In Figure 1, we depict a typical sample path of a marked point process. Let $L(0) = 0$ and

$$L(n) = \sum_{k=0}^{n-1} \ell(k), \quad n = 1, 2, \ldots. \tag{10}$$

Thus, $L(n)$ is the sum of the service requirements of the first n customers. For a marked point process $\Psi = (\tau, \ell)$, we assume in this paper that τ is an increasing sequence, i.e., $\tau \in \mathcal{F}_{inc}$, and that ℓ is a nonnegative integer-valued sequence. Under such an assumption for ℓ, the sequence $L = \{L(n), n = 0, 1, 2, \ldots\}$ is an increasing integer-valued sequence with $L(0) = 0$.

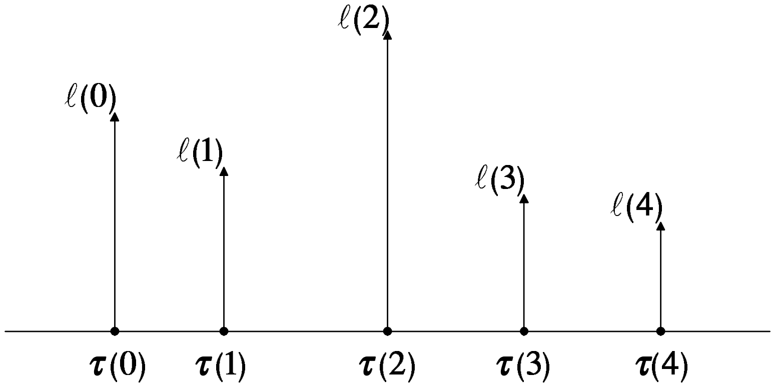

Figure 1: A marked point process

Definition 3.1 (Traffic characterization) *A marked point process $\Psi = (\tau, \ell)$ is said to be g-regular for some $g \in \mathcal{F}_{inc}^0$ if for all $m \le n$ there holds $\tau(n) - \tau(m) \ge g(L(n) - L(m))$.*

As $g(0) = 0$, this characterization can be rewritten as

$$\tau(n) = \max_{0 \le m \le n} [\tau(m) + g(L(n) - L(m))], \tag{11}$$

or simply,

$$\tau = \tau \star_L g. \tag{12}$$

It is straightforward to verify the following two facts.

(i) (Shift invariance) If a marked point process $\Psi = (\tau, \ell)$ is g-regular, then the marked point process $\Psi^1 = (\tau^1, \ell)$ with $\tau^1(n) = \tau(n) + d$ for some fixed d is also g-regular.

(ii) (Monotonicity) If a marked point process $\Psi = (\tau, \ell)$ is g_1-regular and $g_1 \geq g_2$ for some $g_2 \in \mathcal{F}^0_{inc}$, then Ψ is also g_2-regular.

Note that the traffic characterization in Definition 3.1 provides a lower bound on the distance between two arrivals, while Cruz's traffic characterization in (1) provides an upper bound on the number of arrivals within an interval. In the following, we show that these two traffic characterzations are related by the inversion formula.

For an increasing function $g : \mathcal{R}^+ \cup \{0\} \mapsto \mathcal{R}$, define its upper and lower inverse functions as follows:

$$g_U^{-1}(t) = \inf\{s \geq 0 : g(s) > t\},$$
$$g_L^{-1}(t) = \inf\{s \geq 0 : g(s) \geq t\}.$$

Note that both the upper and lower inverse functions are still increasing. Moreover, for all $\tau \in \mathcal{R}^+ \cup \{0\}$

$$g_L^{-1}(g(\tau)) \leq \tau \leq g_U^{-1}(g(\tau)). \tag{13}$$

In particular, if g is *strictly* increasing and continuous, then we have $g_L^{-1}(t) = g_U^{-1}(t)$ and we simply write the inverse function of g as g^{-1}.

Lemma 3.2 **(Inversion formula)** *Consider a marked point process $\Psi = (\tau, \ell)$. Let ℓ_{max} be the maximum service requirement and $V(t)$ be the cumulative service requirements (workload) by time t, i.e.,*

$$\ell_{\max} = \sup_{n \geq 0} \ell(n) \tag{14}$$
$$V(t) = \sum_{\{n:\tau(n) \leq t\}} \ell(n). \tag{15}$$

(i) *If Ψ is g-regular, then for all $s \leq t$*

$$V(t) - V(s) \leq g_U^{-1}(t - s) + \ell_{max},$$

where

$$g_U^{-1}(t) = \inf\{s \geq 0 : \tilde{g}(s) > t\} \tag{16}$$

is the upper inverse function of any function $\tilde{g} : \mathcal{R}^+ \cup \{0\} \mapsto \mathcal{R}$ *with* $\tilde{g}(n) = g(n)$ *for all* $n = 0, 1, 2, \ldots$.

(ii) *If* $V(t) - V(s) \leq f(t-s)$ *for all* $s \leq t$ *and some nonnegative and increasing function* $f(t)$, $t \geq 0$, *then* Ψ *is* f_L^{-1}*-regular, where*

$$f_L^{-1}(n) = \inf\{t \geq 0 : f(t) \geq n\} \tag{17}$$

is the lower inverse function of f.

The inversion formula shows that the characterization in Definition 3.1 is weaker than Cruz's characterization [9] for $V(t)$. Also, it keeps the setting in discrete time and thus avoids the continuity problem for $V(t)$ in Cruz's characterization.

Proof (i) Suppose that $\tau(n-1) \leq t < \tau(n)$ and $\tau(m-1) \leq s < \tau(m)$ for some $m < n$ (the case with $m = n$ is trivial). Then

$$V(t) - V(s) = L(n) - L(m) \leq L(n-1) - L(m) + \ell_{max}. \tag{18}$$

As we assume that Ψ is g-regular,

$$\tau(n-1) - \tau(m) \geq g(L(n-1) - L(m)) = \tilde{g}(L(n-1) - L(m))$$

for any function $\tilde{g} : \mathcal{R}^+ \cup \{0\} \mapsto \mathcal{R}$ with $\tilde{g}(n) = g(n)$ for all $n = 0, 1, 2, \ldots$. Taking the upper inverse function yields

$$L(n-1) - L(m) \leq g_U^{-1}\Big(\tilde{g}(L(n-1) - L(m))\Big) \leq g_U^{-1}(\tau(n-1) - \tau(m)).$$

From (18), it then follows that

$$V(t) - V(s) \leq g_U^{-1}(\tau(n-1) - \tau(m)) + \ell_{max} \leq g_U^{-1}(t-s) + \ell_{max},$$

as $\tau(n-1) \leq t$ and $\tau(m) \geq s$.

(ii) Take $t = \tau(n)$ and $s = \tau(m) - \epsilon$ for some arbitrary $\epsilon > 0$. It then follows from the assumption for $V(t)$ that

$$f(\tau(n) - \tau(m) + \epsilon) \geq V(\tau(n)) - V(\tau(m) - \epsilon) \geq L(n) - L(m),$$

as $V(\tau(n)) \geq L(n)$ and $V(\tau(m)-\epsilon) \leq L(m)$. Taking the lower inverse function yields

$$\begin{aligned}\tau(n)-\tau(m)+\epsilon &\geq f_L^{-1}\Big(f(\tau(n)-\tau(m)+\epsilon)\Big)\\ &\geq f_L^{-1}(L(n)-L(m)).\end{aligned}$$

Letting $\epsilon \to 0$ completes the proof. ■

3.2 Minimal g-regulator

In this section, we show that one can construct a traffic regulator such that its output is g-regular, provided that the sequence g is superadditive. The construction is analogous to the maximal regulator in [5, 6].

Theorem 3.3 (Minimal regulator) *Suppose that $g \in \mathcal{F}_{inc}^0$ and g is superadditive. Consider a marked point process $\Psi = (\tau, \ell)$. Let $\tau^1 = \tau \star_L g$, i.e.,*

$$\tau^1(n) = \max_{0 \leq m \leq n} [\tau(m) + g(L(n) - L(m))],$$

for all n. Construct the marked point process $\Psi^1 = (\tau^1, \ell)$.

(i) *(Traffic regulation) Ψ^1 is g-regular.*

(ii) *(Optimality) For any g-regular marked point process $\tilde{\Psi} = (\tilde{\tau}, \ell)$ with $\tilde{\tau} \geq \tau$, we have $\tilde{\tau} \geq \tau^1$.*

(iii) *(Conformity) Ψ is g-regular if and only if $\Psi^1 = \Psi$.*

The construction of Ψ^1 is called the minimal g-regulator *(for the input Ψ) in the paper (see Figure 2).*

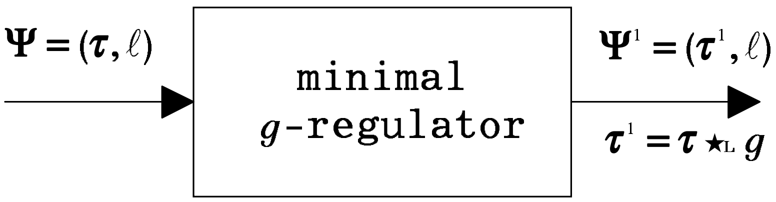

Figure 2: The minimal g-regulator

Proof (i) In view of (12), it suffices to show that $\tau^1 \star_L g = \tau^1$. Note that

$$
\begin{aligned}
&(\tau^1 \star_L g)(n) = ((\tau \star_L g) \star_L g)(n) \\
&= \max_{0\le m\le n} \Big[\max_{0\le u\le m} [\tau(u) + g(L(m) - L(u))] + g(L(n) - L(m))\Big] \\
&= \max_{0\le u\le n} \Big[\tau(u) + \max_{u\le m\le n} [g(L(m) - L(u)) + g(L(n) - L(m))]\Big].
\end{aligned}
$$

As g is superadditive and $g(0) = 0$,

$$\max_{u\le m\le n} [g(L(m) - L(u)) + g(L(n) - L(m))] = g(L(n) - L(u)).$$

Thus,

$$
\begin{aligned}
&(\tau^1 \star_L g)(n) = \max_{0\le u\le n} [\tau(u) + g(L(n) - L(u))] \\
&= (\tau \star_L g)(n) = \tau^1(n).
\end{aligned}
$$

(ii) As $\tilde{\Psi}$ is assumed to be g-regular, we have from (12) that $\tilde{\tau} = \tilde{\tau} \star_L g$. From the assumption $\tilde{\tau} \ge \tau$ and the monotonicity of $\star_L$, it follows that $\tilde{\tau} \ge \tau \star_L g = \tau^1$.

(iii) If $\Psi^1 = \Psi$, it then follows from (i) that Ψ is g-regular. To prove the converse, suppose that Ψ is g-regular. Then $\tau = \tau \star_L g = \tau^1$. Thus, $\Psi = \Psi^1$. ■

As in [5, 6], we note that the condition $\tilde{\tau} \ge \tau$ in Theorem 3.3(ii) is one of the causal conditions as a customer cannot depart before it arrives. Theorem 3.3(i) shows that for any input Ψ, the minimal g-regulator generates an output Ψ^1 that is g-regular. Theorem 3.3(ii) shows that it is the best construction that one can implement if one would like to *minimize* the departure times. This is also the reason that we name it minimal regulator. Finally, Theorem 3.3(iii) shows that if Ψ is already g-regular, then it will pass through the minimal g-regulator without any change.

Example 3.4 ($G/G/1$ queue) Consider feeding a marked point process $\Psi = (\tau, \ell)$ to a $G/G/1$ queue with $\tau(n)$ and $\ell(n)$ being the arrival time and the service requirement of the $n + 1^{th}$ customer. Suppose that the service rate of the $G/G/1$ queue is r. Let $\Psi^1 = (\tau^1, \ell)$ with $\tau^1(n)$ being the time that the $n + 1^{th}$ customer starts being served. It is well known that such a queue is governed by the Lindley equation as follows:

$$
\begin{aligned}
&\tau^1(0) = \tau(0), \\
&\tau^1(n) = \max[\tau(n), \tau^1(n-1) + \frac{\ell(n-1)}{r}], \quad n \ge 1.
\end{aligned}
\tag{19}
$$

Expanding recursively yields

$$\tau^1(n) = \max_{0 \le m \le n} [\tau(m) + \frac{1}{r}(L(n) - L(m))]. \tag{20}$$

Thus, Ψ^1 is the output from the minimal g-regulator with $g(n) = n/r$ for all n.

Example 3.5 (Shifted rate regulator) Example 3.4 shows that Lindley's recursion in (19) can be used for implementing the minimal g-regulator with $g(n) = n/r$ for all n. Now we show the same recursion can also be used for implementing the minimal g-regulator with $g(n) = \frac{1}{r}(n-d)^+$, where $(x)^+ = \max[0, x]$. Such a regulator is called the shifted rate regulator with rate r and delay d, and it is the counterpart of the leaky bucket in the min-plus algebra.

Let $\Psi = (\tau, \ell)$ and $\Psi^1 = (\tau^1, \ell)$ be the input and the output from such a regulator. Note that $g(n) = \frac{1}{r}(n-d)^+$ is superadditive. Thus, from the construction of the minimal g-regulator in Theorem 3.3, it follows that

$$\begin{aligned}
\tau^1(n) &= \max_{0 \le m \le n} [\tau(m) + \frac{1}{r}(L(n) - L(m) - d)^+] \\
&= \max_{0 \le m \le n} \Big[\tau(m) + \max[\frac{1}{r}(L(n) - L(m) - d), 0]\Big] \\
&= \max\Big[\max_{0 \le m \le n} [\tau(m)], \max_{0 \le m \le n} [\tau(m) + \frac{1}{r}(L(n) - L(m))] - \frac{d}{r}\Big] \\
&= \max[\tau(n), (\tau \star_L g_2)(n) - \frac{d}{r}],
\end{aligned}$$

where $g_2(n) = n/r$ for all n. Let $\tau^2 = \tau \star_L g_2$. In view of the $G/G/1$ queue in Example 3.4, the shifted rate regulator with rate r and delay d can be implemented by the following recursive equations:

$$\tau^2(0) = \tau(0), \tag{21}$$

$$\tau^2(n) = \max[\tau(n), \tau^2(n-1) + \frac{\ell(n-1)}{r}], \tag{22}$$

$$\tau^1(n) = \max[\tau(n), \tau^2(n) - \frac{d}{r}]. \tag{23}$$

3.3 Minimal g-regulators in parallel

A sequence g is called star-shaped (see e.g., Marshall and Olkin [23]) if

$$\frac{g(m)}{m} \le \frac{g(n)}{n}, \quad m \le n.$$

A star-shaped sequence g is also superadditive as

$$g(m) + g(n-m) \le \frac{m}{n} g(n) + \frac{n-m}{n} g(n) = g(n), \quad m \le n.$$

Also, the family of star-shaped sequences are max-closed, i.e., $\max[g_1(n), g_2(n)]$ is also star-shaped if both g_1 and g_2 are star-shaped. In the following theorem, we show that one can construct a composite minimal regulator by combining simple minimal regulators in parallel provided that the constraint sequences are star-shaped.

Theorem 3.6 (Filter bank summation) *Consider a marked point process $\Psi = (\tau, \ell)$. Let $\Psi^1 = (\tau^1, \ell)$ (resp. $\Psi^2 = (\tau^2, \ell)$) be the output from the minimal g_1-regulator (resp. the minimal g_2-regulator) for Ψ. The output from the "filter bank summation", denoted by $\Psi^3 = (\tau^3, \ell)$, is constructed from $\tau^3 = \tau^1 \oplus \tau^2$. If both g_1 and g_2 are* star-shaped*, then the "filter bank summation" of the minimal g_1-regulator for Ψ and the minimal g_2-regulator for Ψ is the minimal g-regulator for Ψ, where $g = g_1 \oplus g_2$.*

Proof As we assume that Ψ^1 is the output from the minimal g_1-regulator for Ψ,

$$\tau^1 = \tau \star_L g_1.$$

Similarly,

$$\tau^2 = \tau \star_L g_2.$$

It then follows from the distributive property that

$$\tau^3 = \tau^1 \oplus \tau^2 = (\tau \star_L g_1) \oplus (\tau \star_L g_2) = \tau \star_L (g_1 \oplus g_2).$$

As we assume that both g_1 and g_2 are star-shaped, $g_1 \oplus g_2$ is also star-shaped and thus superadditive. It then follows from Theorem 3.3(i) that the the "filter bank summation" of the minimal g_1-regulator and the minimal g_2-regulator is the minimal g-regulator. ■

Example 3.7 (CXPL regulators) Note that $g(n) = \frac{1}{r}(n-d)^+$ is star-shaped. From Theorem 3.6, one can implement the minimal g-regulator with

$$g(n) = \max_{1 \le k \le K} [\frac{1}{r_k}(n - d_k)^+]$$

by combining K shifted rate regulators with rates r_k and delays d_k, $k = 1, \ldots, K$. Such a regulator is called a convex piecewise linear (CXPL) regulator in this paper. Let $\Psi = (\tau, \ell)$ and $\Psi^1 = (\tau^1, \ell)$ be the input and the output from such a regulator. In view of Example 3.5, a CXPL regulator can be constructed by the following recursive equations:

$$\tau_k^2(0) = \tau(0), \quad k = 1, \ldots, K, \tag{24}$$

$$\tau_k^2(n) = \max[\tau(n), \tau_k^2(n-1) + \frac{\ell(n-1)}{r_k}] \quad k = 1, \ldots, K, \tag{25}$$

$$\tau^1(n) = \max\Big[\tau(n), \max_{1 \le k \le K} [\tau_k^2(n) - \frac{d_k}{r_k}]\Big]. \tag{26}$$

Remark 3.8 We note that not every result in [5] can be extended here. For example, a concatenation of the minimal g_1-regulator and the minimal g_2-regulator is not the minimal $g_1 \star g_2$-regulator in general. This is because we only have an inequality in Lemma 2.1. However, if both g_1 and g_2 are star-shaped, then one can easily show that a concatenation of the minimal g_1-regulator and the minimal g_2-regulator is the minimal $g_1 \oplus g_2$-regulator.

3.4 Superposition of g-regular traffic

In this section, we show how one obtains traffic characterization for superpositioning (multiplexing) g-regular traffic. Our approach is based on the inversion formula in Lemma 3.2.

Corollary 3.9 (Superposition) *Consider K marked point process $\Psi_k = (\tau_k, \ell_k)$, $k = 1, \ldots, K$. Let $\Psi = (\tau, \ell)$ be the superposition of these K marked point processes. If Ψ_k is g_k-regular, then Ψ is g-regular with*

$$g(n) = \inf\{t \ge 0 : \sum_{k=1}^{K} g_{U,k}^{-1}(t) + \ell_{k,max} \ge n\},$$

where $g_{U,k}^{-1}$ is the upper inverse function of g_k and $\ell_{k,max}$ is the maximum service requirement of ℓ_k as defined in Lemma 3.2.

Proof Let $V_k(t)$ be the cumulative service requirements by time t for Ψ_k, and $V(t)$ be the cumulative service requirements by time t for Ψ. From Lemma 3.2, it follows that for all $s \le t$, $k = 1, \ldots, K$,

$$V_k(t) - V_k(s) \le g_{U,k}^{-1}(t-s) + \ell_{k,\max}.$$

Since Ψ is the superposition of these K marked point processes, $V(t) = \sum_{k=1}^K V_k(t)$. Thus,

$$V(t) - V(s) \le \sum_{k=1}^{K} g_{U,k}^{-1}(t-s) + \ell_{k,\max}.$$

Applying Lemma 3.2(ii) completes the proof. ■

3.5 Segmentation and reassembly

In a telecommunication network, sometimes it is necessary to transmit variable length packets by smaller packets (cells) with the same size. The process that separates variable length packets into smaller packets with the same size is called *segmentation*. On the other hand, the process that reassembles smaller packets back to the original packets is called *reassembly*.

Consider a marked point process $\Psi = (\tau, \ell)$. Assume that $\ell(n) > 0$ for all n. Its output after segmentation is the marked point process, $\Psi^s = (\tau^s, \ell^s)$, where

$$\ell^s(n) = 1, \quad \text{for all } n, \tag{27}$$

$$\tau^s(n) = \tau(m), \quad L(m) \leq n < L(m+1), \tag{28}$$

i.e., cells from the same packet have the same arrival times. On the other hand, for a segmented marked point process $\Psi^s = (\tau^s, \ell^s)$, its output after reassembly is the marked point process $\Psi^r = (\tau^r, \ell^r)$, where

$$\ell^r(n) = \ell(n), \quad \text{for all } n, \tag{29}$$

$$\tau^r(n) = \tau^s(L(n+1) - 1), \quad \text{for all } n, \tag{30}$$

i.e., the time to reassemble a packet is the arrival time of the last cell in that packet.

Lemma 3.10 (Segmentation and reassembly) *Assume that $\ell(n) > 0$ for all n. Let $\ell_{max} = \sup_{n\geq 0} \ell(n)$.*

(i) *If Ψ is g-regular, then Ψ^s is g^s-regular, where $g^s(t) = g((t - \ell_{max} + 1)^+)$.*

(ii) *If Ψ^s is g-regular, then Ψ^r is g^r-regular, where $g^r(t) = g((t - \ell_{max} + 1)^+)$.*

Proof (i) For $k_2 < k_1$, assume that $L(n) \leq k_1 < L(n+1)$ and $L(m) \leq k_2 < L(m+1)$ for some $m \leq n$. Thus, $k_1 \leq L(n) + \ell_{max} - 1$. It then follows from (28) and g-regularity for Ψ that

$$\begin{aligned} \tau^s(k_1) - \tau^s(k_2) &= \tau(n) - \tau(m) \geq g(L(n) - L(m)) \\ &\geq g((k_1 - (\ell_{max} - 1) - k_2)^+). \end{aligned}$$

This shows that Ψ^s is g^s-regular.

(ii) Note from (30) and g-regularity for Ψ^s that

$$\begin{aligned}&\tau^r(n) - \tau^r(m) = \tau^s(L(n+1)-1) - \tau^s(L(m+1)-1)\\&\geq g((L(n+1)-1)-(L(m+1)-1)) = g(L(n+1)-L(m+1)).\end{aligned}$$

Since $L(n+1) \geq L(n)+1$ and $L(m+1) \leq L(m)+\ell_{max}$, we have

$$\tau^r(n) - \tau^r(m) \geq g((L(n)-L(m)-(\ell_{max}-1))^+).$$

This shows that Ψ^r is g^r-regular. ■

We note that the results in Lemma 3.10 are without segmentation and reassembly delay. The case with segmentation and reassembly delay can be treated by the $\mathbf{O}_d$-server discussed in the next section.

4 Service guarantees for marked point processes

4.1 g-server

In this section, we extend the concept of service curve in [25, 5, 1, 21] to guarantee deterministic quality-of-service for marked point processes.

Definition 4.1 (g-server) *A server is called a g-server ($g \in \mathcal{F}_{inc}$) for an input marked point process $\Psi = (\tau, \ell)$ if its output marked point process $\Psi^1 = (\tau^1, \ell)$ satisfies $\tau^1 \leq \tau \star_L g$, i.e.,*

$$\tau^1(n) \leq \max_{0 \leq m \leq n} [\tau(m) + g(L(n)-L(m))] \tag{31}$$

for all n (see Figure 3). If the inequality in (31) is satisfied for all input marked point processes, then we say the g-server is universal. *If the inequality is an equality, we say the g-server is* exact.

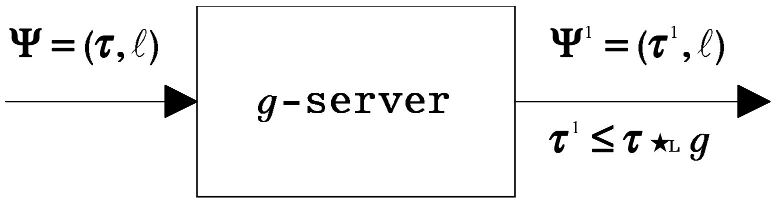

Figure 3: A g-server

Clearly, the minimal g-regulator is a *universal* and *exact* g-server. Also, we note that if a server is a g_1-server for a marked point process $\Psi = (\tau, \ell)$ and $g_1 \leq g_2$ for some $g_2 \in \mathcal{F}_{inc}$, then the server is also a g_2-server for Ψ.

Example 4.2 (Segmentation and reassembly) Consider transmitting a marked point process $\Psi = (\tau, \ell)$ via segmentation and reassembly in Section 3.5 (see Figure 4). Assume that $\ell(n) > 0$ for all n. Let $\Psi^s = (\tau^s, \ell^s)$ be the output of Ψ after segmentation in (27) and (28). The marked point process Ψ^s is then fed into a g-server. Let $\Psi^{s,1} = (\tau^{s,1}, \ell^s)$ be the output from the g-server. From Definition 4.1 and $\ell^s(n) = 1$ for all n in (27), it follows that

$$\tau^{s,1}(n) \leq \max_{0 \leq m \leq n} [\tau^s(m) + g(n-m)]. \tag{32}$$

The marked point process $\Psi^{s,1}$ is then reassembled via (29) and (30). Let $\Psi^1 = (\tau^1, \ell)$ be the output after reassembly. From (30) and (32), we have

$$\tau^1(n) = \tau^{s,1}(L(n+1)-1) \leq \tau^s(m^*) + g(L(n+1) - 1 - m^*),$$

for some $0 \leq m^* < L(n+1)$. Now suppose that $L(m) \leq m^* < L(m+1)$ for some $0 \leq m \leq n$. In conjunction with (28),

$$\begin{aligned}
\tau^1(n) &\leq \tau(m) + g(L(n+1) - 1 - m^*) \\
&\leq \tau(m) + g(L(n) + \ell_{max} - 1 - L(m)) \\
&\leq \max_{0 \leq k \leq n} [\tau(k) + g(L(n) - L(k) + \ell_{max} - 1)],
\end{aligned}$$

where $\ell_{max} = \sup_{n \geq 0} \ell(n)$ is the maximum service requirement. This shows that Ψ^1 is the output from a g_1-server for Ψ, where $g_1(t) = g(t + \ell_{max} - 1)$.

In the following theorem, we derive performance bounds for g-servers.

Theorem 4.3 (Performance bounds) *Consider a g_2-server for a marked point process $\Psi = (\tau, \ell)$. Let $\Psi^1 = (\tau^1, \ell)$ be the output. Also, let $d = \sup_{n \geq 0}[\tau^1(n) - \tau(n)]^+$ be the maximum delay at the server. Suppose that Ψ is g_1-regular.*

(i) *(Maximum delay) $d \leq \sup_{n \geq 0}[g_2(n) - g_1(n)]^+$.*

(ii) *(Maximum queued service requirements) The total amount of service requirements queued at the server is bounded above by $g_{U,1}^{-1}(d) + \ell_{max}$, where d is the maximum delay in (i), $g_{U,1}^{-1}$ is the upper inverse function of g_1 defined in (16), and ℓ_{max} is the maximum service requirement defined in (14).*

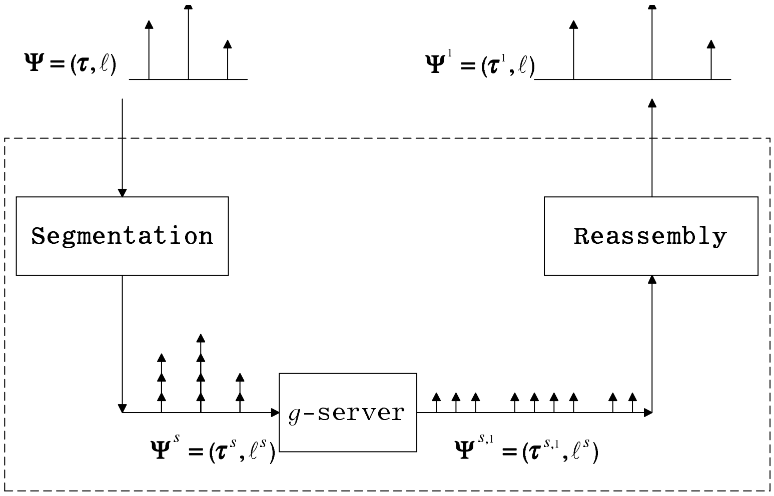

Figure 4: Segmentation and reassembly

(iii) *(Output characterization) If $\tau^1 \geq \tau$, then Ψ^1 is g_3-regular, where $g_3(0) = 0$ and*

$$g_3(n) = \inf_{m \geq 0} [g_1(m+n) - g_2(m)]^+, \quad n > 0.$$

Proof (i) As we assume that the server is a g_2-server for Ψ and that Ψ is g_1-regular, we have

$$\begin{aligned}
&\tau^1(n) - \tau(n) \\
&\leq \max_{0 \leq m \leq n} [\tau(m) + g_2(L(n) - L(m))] - \tau(n) \\
&= \max_{0 \leq m \leq n} [g_2(L(n) - L(m)) - (\tau(n) - \tau(m))] \\
&\leq \max_{0 \leq m \leq n} [g_2(L(n) - L(m)) - g_1(L(n) - L(m))] \\
&\leq \max_{0 \leq m \leq n} \left[\sup_{k \geq 0} [g_2(k) - g_1(k)]\right] \\
&= \sup_{k \geq 0} [g_2(k) - g_1(k)].
\end{aligned}$$

(ii) Let $q(t)$ be the total amount of service requirements queued at the server at time t. Also let $V(t)$ and $V^1(t)$ be the cumulative service requirements by time t for the input Ψ and the output Ψ^1, respectively. From Theorem 4.3(i), the maximum delay is bounded above by d. This implies

$$V^1(t) = \sum_{\{n:\tau^1(n) \leq t\}} \ell(n) \geq \sum_{\{n:\tau(n)+d \leq t\}} \ell(n) = V(t-d).$$

Thus,

$$q(t) = V(t) - V^1(t) \leq V(t) - V(t-d) \leq g_{U,1}^{-1}(d) + \ell_{\max},$$

where we apply Lemma 3.2(i) in the last inequality.

(iii) As we assume that $\tau^1 \geq \tau$ and that Ψ^1 is the output from the g_2-server for Ψ, we have for $m < n$

$$\begin{aligned}
\tau^1(n) - \tau^1(m) &\geq \tau(n) - \max_{0\leq k\leq m}[\tau(k) + g_2(L(m) - L(k))] \\
&= \min_{0\leq k\leq m}[\tau(n) - \tau(k) - g_2(L(m) - L(k))] \\
&\geq \min_{0\leq k\leq m}[g_1(L(n) - L(k)) - g_2(L(m) - L(k))],
\end{aligned}$$

where we use the assumption that Ψ is g_1-regular in the last inequality. Using the change of variable $v = L(m) - L(k)$ yields

$$\tau^1(n) - \tau^1(m) \geq \min_{0\leq v\leq L(m)}[g_1(L(n) - L(m) + v) - g_2(v)]. \tag{33}$$

As τ^1 is increasing, $\tau^1(n) - \tau^1(m) \geq 0$ for $m < n$. In conjunction with (33),

$$\begin{aligned}
\tau^1(n) - \tau^1(m) &\geq \max\Big[0, \min_{0\leq v\leq L(m)}[g_1(L(n) - L(m) + v) - g_2(v)]\Big] \\
&= \min_{0\leq v\leq L(m)}[g_1(L(n) - L(m) + v) - g_2(v)]^+ \\
&\geq \inf_{v\geq 0}[g_1(L(n) - L(m) + v) - g_2(v)]^+ = g_3(L(n) - L(m)).
\end{aligned}$$

■

Lemma 4.4 ($\mathbf{O}_d$-server) *Let $\mathbf{O}_d$ be the sequence with $\mathbf{O}_d(n) = d$ for all $n = 0, 1, 2 \ldots$.*

(i) *A server guarantees maximum delay d for a marked point process $\Psi = (\tau, \ell)$ if and only if it is an $\mathbf{O}_d$-server for Ψ.*

(ii) *Suppose that Ψ is g-regular. Then a server is an $\mathbf{O}_d$-server for Ψ if and only if the server is $g \star \mathbf{O}_d$-server for Ψ.*

Proof (i) Note that for all n

$$\tau(n) + d = \max_{0\leq m\leq n}[\tau(m) + d] = \max_{0\leq m\leq n}[\tau(m) + \mathbf{O}_d(L(m) - L(n))] = (\tau \star_L \mathbf{O}_d)(n).$$

Let $\Psi^1 = (\tau^1, \ell)$ be the output from the server. It is then clear that $\tau^1(n) \leq \tau(n) + d$ if and only if $\tau^1(n) \leq (\tau \star_L \mathbf{O}_d)(n)$.

(ii) Since $g \in \mathcal{F}_{inc}^{0}$,

$$(g \star \mathbf{O}_d)(n) = g(n) + d \geq g(0) + d = d = \mathbf{O}_d(n).$$

Thus, if the server is an $\mathbf{O}_d$-server for Ψ, it is also a $g \star \mathbf{O}_d$-server for Ψ.

Now suppose that the server is a $g \star \mathbf{O}_d$-server for Ψ. Since Ψ is assumed to be g-regular, it follows from Theorem 4.3(i) that the maximum delay is bounded above by d. Thus, we have from (i) that the server is an $\mathbf{O}_d$-server for Ψ. ■

4.2 g-servers in tandem

In this section, we discuss properties for g-servers in tandem.

Theorem 4.5 (Concatenation) *A concatenation of a g_1-server for a marked point process $\Psi = (\tau, \ell)$ and a g_2-server for the output from the g_1-server is a g-server for Ψ, where $g = g_1 \star g_2$ (see Figure 5).*

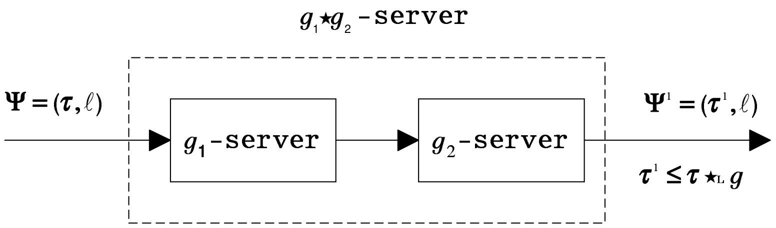

Figure 5: A concatenation of two servers

Proof Let $\Psi^1 = (\tau^1, \ell)$ be the output from the g_1-server and $\Psi^2 = (\tau^2, \ell)$ be the output from the g_2-server. It then follows that

$$\tau^1 \leq \tau \star_L g_1,$$
$$\tau^2 \leq \tau^1 \star_L g_2.$$

From the monotonicity of $\star_L$ and Lemma 2.1, we have

$$\tau^2 \leq (\tau \star_L g_1) \star_L g_2 \leq \tau \star_L (g_1 \star g_2).$$

Thus, Ψ^2 is the output from a $g_1 \star g_2$-server for Ψ. ■

Example 4.6 ($G/G/1$ queue) Continue from Example 3.4. Let $\Psi^2 = (\tau^2, \ell)$ with $\tau^2(n)$ being the departure time of the $n+1^{th}$ customer. As $\tau^1(n)$ is the time that the $n+1^{th}$ customer starts being served, we have

$$\tau^2(n) = \tau^1(n) + \frac{\ell(n)}{r} \leq \tau^1(n) + d = (\tau^1 \star_L \mathbf{O}_d)(n),$$

where $d = \ell_{max}/r$ and ℓ_{max} is the maximum service requirement of all the customers. Thus, a $G/G/1$ queue with rate r can be viewed as a concatenation of a g_1-server and an $\mathbf{O}_d$-server with $g_1(n) = n/r$ and $d = \ell_{max}/r$. From Theorem 4.5, it is a g_2-server with $g_2(n) = (n + \ell_{max})/r$.

Example 4.7 ($G/G/1$ queue with vacation) Continue from Example 4.6. Suppose that the server takes a vacation every time when the queue is empty. When a vacation ends, it will take another vacation if the queue is still empty. Otherwise, it starts to serve the first customer in the queue. Suppose that the vacation time is bounded above by v_{max}. From a standard sample path argument, it follows that the departure time of a customer in a $G/G/1$ queue with vacation is not later than the sum of the departure time of that customer in the corresponding $G/G/1$ queue and the maximum vacation time v_{max}. Thus, a $G/G/1$ queue with vacation can be viewed as a concatenation of a g_2-server and an $\mathbf{O}_d$-server with $g_2(n) = (n + \ell_{max})/r$ and $d = v_{max}$. Once again, we have from Theorem 4.5 that a $G/G/1$ queue with bounded vacation times is a g_3-server with

$$g_3(n) = \frac{n + \ell_{max}}{r} + v_{max}.$$

Example 4.8 (Time varying rate server) Continue from the $G/G/1$ queue in Example 4.6. Now suppose that the service rate is time varying. Let $r(t)$ be the service rate at time t and $R(t)$ be the cumulative service efforts up to time t, i.e.,

$$R(t) = \int_{-\infty}^{t} r(s) ds. \tag{34}$$

Assume that

$$R(t) - R(s) \geq f(t-s), \quad s \leq t \tag{35}$$

for some increasing function f with $f(0) = 0$. As in Example 3.4 and Example 4.6, let $\Psi = (\tau, \ell)$ be the input, $\Psi^1 = (\tau^1, \ell)$ be the output when customers start their services, and $\Psi^2 = (\tau^2, \ell)$ be the output when customers complete their services. Let $\ell_{max} = \sup_{n \geq 0} \ell(n)$ be the maximum service requirement of all the customers. We show in the following that Ψ^1 is the output from a f_U^{-1}-server for Ψ, and Ψ^2 is the output from a $f_U^{-1} \star \mathbf{O}_d$-server for Ψ, where $f_U^{-1}(n) = \inf\{t \geq 0 : f(t) > n\}$ is the upper inverse function of f and $d = f_U^{-1}(\ell_{max})$.

Let $\tilde{\ell}(n)$ be the service time of the $n+1^{th}$ customer at the time varying rate server. Then

$$\tilde{\ell}(n) = \inf\{t \geq 0 : R(\tau^1(n) + t) - R(\tau^1(n)) > \ell(n)\}. \tag{36}$$

As $R(t)$ in (34) is continuous, we have from (36) that

$$R(\tau^1(n) + \tilde{\ell}(n)) = R(\tau^1(n)) + \ell(n). \tag{37}$$

It then follows from the Lindley equation $\tau^1(n) = \max[\tau(n), \tau^1(n-1) + \tilde{\ell}(n-1)]$ that

$$R(\tau^1(n)) = \max[R(\tau(n)), R(\tau^1(n-1)) + \ell(n-1)]. \tag{38}$$

Equivalently,

$$R(\tau^1(n)) = \max_{0 \leq m \leq n}[R(\tau(m)) + L(n) - L(m)].$$

as $\tau^1(0) = \tau(0)$. This implies that for some $0 \leq m^* \leq n$

$$L(n) - L(m^*) = R(\tau^1(n)) - R(\tau(m^*)).$$

Since we assume that $R(t) - R(s) \geq f(t-s)$ (in fact we only need a weaker assumption that $R(\tau^1(n)) - R(s) \geq f(\tau^1(n) - s)$),

$$L(n) - L(m^*) \geq f(\tau^1(n) - \tau(m^*)). \tag{39}$$

Taking the upper inverse function on both sides of (39) yields

$$\tau^1(n) \leq \tau(m^*) + f_U^{-1}(L(n) - L(m^*)) \leq \max_{0 \leq m \leq n}[\tau(m) + f_U^{-1}(L(n) - L(m)],$$

as $0 \leq m^* \leq n$. This shows that Ψ^1 is the output from an f_U^{-1}-server for Ψ.

Let $\tilde{\ell}_{max} = \sup_{n \geq 0} \tilde{\ell}(n)$. As in Example 4.6, Ψ^2 is an $\mathbf{O}_d$-server for Ψ^1, where $d = \tilde{\ell}_{max}$. From (36) and the assumption that $R(t) - R(s) \geq f(t-s)$, it follows that

$$\tilde{\ell}_{max} \leq f_U^{-1}(\ell_{max}).$$

Thus, Ψ^2 is the output from a $f_U^{-1} \star \mathbf{O}_d$-server for Ψ, where $d = f_U^{-1}(\ell_{max})$.

Example 4.9 (Guaranteed rate server) In Goyal and Vin [20], a server is called a guaranteed rate (GR) server with rate r if the departure time of a customer is not later than the sum of the departure time from a $G/G/1$ queue with rate r and a constant β. Clearly, the vacation model in Example 4.7 can be modelled as a GR server. Other examples of GR servers, including Virtual Clock (VC) [27] and Packetized Generalized Processor Sharing (PGPS) [24], can be found in [20]. From Example 4.6, a GR server with the rate r and the constant β can be viewed as a concatenation of a g_2-server and an $\mathbf{O}_\beta$-server with $g_2(n) = (n + \ell_{max})/r$. Thus, we have from Theorem 4.5 that a concatenation of K GR servers with the rates r_k and the constants β_k, $k = 1, \ldots, K$, is a $g \star \mathbf{O}_d$-server with $g(n) = n/\min_{1\le k\le K}[r_k]$ and $d = \sum_{k=1}^{K} \ell_{max}/r_k + \beta_k$.

Example 4.10 (Shifted superadditive server) A $\tilde{g}$-server is called a shifted superadditive server if $\tilde{g} = \mathbf{O}_d \star g$ for some $g \in \mathcal{F}_{inc}^0$ that is superadditive. It is shown in Lemma 4.4 that an $\mathbf{O}_d \star g$-server for a g-regular input guarantees the maximum delay d. The condition is necessary and sufficient. Thus, for a g-regular input, one has to implement an $\mathbf{O}_d \star g$-server in order to guarantee maximum delay d.

Now suppose that one implements a concatenation of two shifted superadditive servers: the first one is an $\mathbf{O}_{d_1} \star g_1$-server and the second one is an $\mathbf{O}_{d_2} \star g_2$-server. From Theorem 4.5, this is a g-server with

$$g = (\mathbf{O}_{d_1} \star g_1) \star (\mathbf{O}_{d_2} \star g_2) = \mathbf{O}_{d_1+d_2} \star g_1 \star g_2,$$

as the operator $\star$ is commutative and associative. As discussed in [5], a better implementation is to have the first one as an $\mathbf{O}_{d_1} \star g_1 \star g_2$-server and the second one as an $\mathbf{O}_{d_2} \star g_1 \star g_2$-server. This is due to the facts that $g_1 \star g_2 \geq g_i$, $i = 1$ and 2, and that

$$(\mathbf{O}_{d_1} \star g_1 \star g_2) \star (\mathbf{O}_{d_2} \star g_1 \star g_2) = \mathbf{O}_{d_1+d_2} \star g_1 \star g_2 = g,$$

where we use $g_i \star g_i = g_i$ for superadditive sequences $g_i \in \mathcal{F}_{inc}^0$, $i = 1$ and 2.

In particular, if one implements the $\mathbf{O}_{d_i} \star g_i$ servers by a concatenation of the minimal g_i-regulator and an $\mathbf{O}_{d_i}$-server, $i = 1$ and 2, it is better to have the same minimal $g_1 \star g_2$-regulator at both servers.

4.3 g-servers in parallel

In this section, we discuss properties for g-servers in parallel.

Theorem 4.11 (Filter bank summation) *Consider a marked point process $\Psi = (\tau, \ell)$. Let $\Psi^1 = (\tau^1, \ell)$ (resp. $\Psi^2 = (\tau^2, \ell)$) be the output from a g_1-server (resp. g_2-server) for Ψ. The output from the "filter bank summation", denoted by $\Psi^3 = (\tau^3, \ell)$, is constructed from $\tau^3 = \tau^1 \oplus \tau^2$ (see Figure 6). Then the "filter bank summation" of a g_1-server for Ψ and a g_2-server for Ψ is a g-server for Ψ, where $g = g_1 \oplus g_2$.*

The proof of Theorem 4.11 is similar to that for Theorem 3.6 and thus omitted.

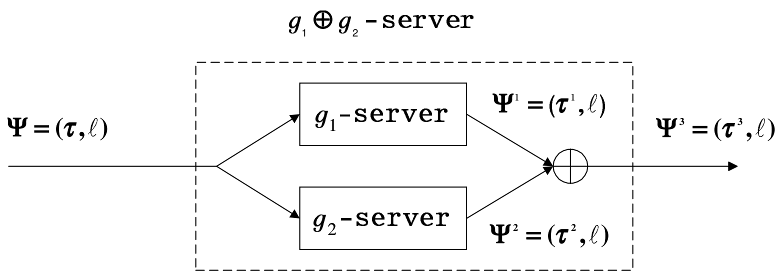

Figure 6: Filter bank summation of two servers

Example 4.12 (Jitter control) Consider feeding a marked point process $\Psi = (\tau, \ell)$ to an $\mathbf{O}_d$-server. Let $\Psi^1 = (\tau^1, \ell)$ be the output from the $\mathbf{O}_d$-server. The maximum jitter between Ψ^1 and Ψ is defined by the maximum difference between the interarrival times of τ^1 and τ, i.e.,

$$j_{max} = \sup_{n \geq 1} |\tau^1(n) - \tau^1(n-1) - (\tau(n) - \tau(n-1))|.$$

Under the causal condition $\tau^1(n) \geq \tau(n)$ for all n, it is easy to see that the maximum jitter j_{max} is bounded above by the maximum delay d. One way to achieve zero jitter is to delay every customer to its maximum delay d. To be precise, let $\Psi^2 = (\tau^2, \ell)$ with $\tau^2(n) = \max[\tau^1(n), \tau(n) + d]$. As Ψ^1 is the output from the $\mathbf{O}_d$-server, we have $\tau^1(n) \leq \tau(n) + d$ and $\tau^2(n) = \tau(n) + d$. Thus, there is no jitter between Ψ^2 and Ψ. In view of this, one can have partial jitter control by setting $\Psi^3 = (\tau^3, \ell)$ with $\tau^3(n) = \max[\tau^1(n), \tau(n) + d_1]$ for some $0 \leq d_1 \leq d$. Equivalently,

$$\tau^3 = \tau^1 \oplus (\tau \star_L \mathbf{O}_{d_1}).$$

Thus, Ψ^3 is the output from the "filter bank summation" of an $\mathbf{O}_d$-server for Ψ and an *exact* $\mathbf{O}_{d_1}$-server for Ψ. It then follows from Theorem 4.11 that Ψ^3 is the output from an $\mathbf{O}_d$-server as $\mathbf{O}_d \oplus \mathbf{O}_{d_1} = \mathbf{O}_d$. On the other hand, one has from the monotonicity of $\oplus$ that $\tau^3 \geq \tau \star_L \mathbf{O}_{d_1}$. Thus, we have for all n

$$\tau(n) + d_1 \leq \tau^3(n) \leq \tau(n) + d. \tag{40}$$

This implies that

$$\tau^3(n) - \tau(n) + \tau(n-1) - \tau^3(n-1) \le d - d_1$$

and that

$$-\tau^3(n) + \tau(n) - \tau(n-1) + \tau^3(n-1) \le d - d_1.$$

The maximum jitter between Ψ^3 and Ψ is then bounded above by $d - d_1$, while the maximum delay is still bounded above by d.

If Ψ is g-regular, then it follows from (40) that for $m < n$

$$\begin{aligned}
&\tau^3(n) - \tau^3(m) \ge \tau(n) + d_1 - (\tau(m) + d) \\
&= \tau(n) - \tau(m) - (d - d_1) \ge g(L(n) - L(m)) - (d - d_1).
\end{aligned}$$

Thus, Ψ^3 is g_3-regular with $g_3(n) = (g(n) - (d - d_1))^+$.

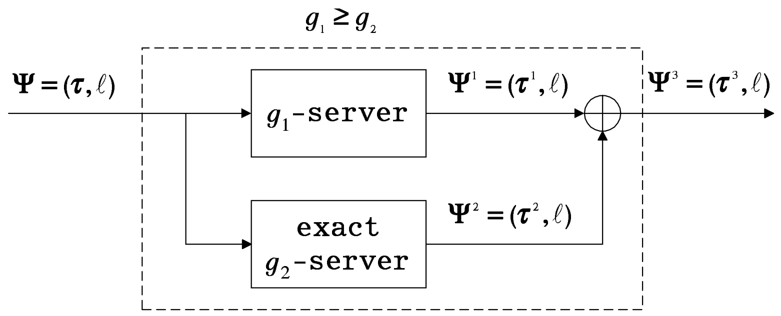

Figure 7: The damper

Example 4.13 (Damper) A more sophisticated jitter control is called a *damper* in [12]. A damper for a marked point process $\Psi = (\tau, \ell)$ is a filter bank summation of a g_1-server and an exact g_2-server with $g_1 \ge g_2$ (see Figure 7). Since $g_1 \oplus g_2 = g_1$, the damper is also a g_1-server. Let $\Psi^1 = (\tau^1, \ell)$, $\Psi^2 = (\tau^2, \ell)$, and $\Psi^3 = (\tau^3, \ell)$ be the outputs from the g_1-server, the g_2-server, and the damper. It then follows from the monotonicity of $\oplus$ that

$$\tau^3 = \tau^1 \oplus \tau^2 \ge \tau^2 = \tau \star_L g_2.$$

Since the damper is also a g_1-server, we have

$$\tau \star_L g_2 \le \tau^3 \le \tau \star_L g_1.$$

Thus, for $m < n$

$$\tau^3(n) - \tau^3(m) \geq \max_{0 \leq u \leq n}[\tau(u) + g_2(L(n) - L(u))] - \max_{0 \leq u \leq m}[\tau(u) + g_1(L(m) - L(u))].$$

Let m^* be the argument that achieves the maximum in $\max_{0 \leq u \leq m}[\tau(u) + g_1(L(m) - L(u))]$, i.e.,

$$\tau(m^*) + g_1(L(m) - L(m^*)) = \max_{0 \leq u \leq m}[\tau(u) + g_1(L(m) - L(u))].$$

Since $0 \leq m^* \leq m$, we have for $m < n$ (cf. the argument for output characterization in Theorem 4.3(iii))

$$\begin{aligned}
&\tau^3(n) - \tau^3(m) \geq \tau(m^*) + g_2(L(n) - L(m^*)) - \tau(m^*) - g_1(L(m) - L(m^*)) \\
&= g_2(L(n) - L(m^*)) - g_1(L(m) - L(m^*)) \\
&\geq \min_{0 \leq u \leq m}[g_2(L(n) - L(u)) - g_1(L(m) - L(u))] \\
&\geq \min_{0 \leq v \leq L(n) - L(m)}[g_2(L(n) - L(m) + v) - g_1(v)] \\
&\geq \inf_{v \geq 0}[g_2(L(n) - L(m) + v) - g_1(v)].
\end{aligned}$$

This implies that the output Ψ^3 from the damper is g_3-regular, where $g_3(0) = 0$ and

$$g_3(n) = \inf_{m \geq 0}[g_2(m + n) - g_1(m)]^+, \quad n \geq 1.$$

Since an exact g-server is equivalent to the filter bank summation of two exact g-servers, the output from an exact g-server is $\tilde{g}$-regular, where $\tilde{g}(0) = 0$ and

$$\tilde{g}(n) = \inf_{m \geq 0}[g(m + n) - g(m)]^+, \quad n \geq 1.$$

In particular, if g is superadditive, then $\tilde{g} = g$ and the damper reduces to the minimal g-regulator.

4.4 g-server with feedback

In this section, we discuss properties for a g-server with feedback.

Theorem 4.14 (Feedback) *Consider a feedback system with the input $\Psi = (\tau, \ell)$ and the output $\Psi^1 = (\tau^1, \ell)$ (see Figure 8). In the feedback system, $\Psi^2 = (\tau^2, \ell)$ is the output from a $g \star \mathbf{I}_w$-server for Ψ^1, and the feedback system is joined by the following equation*

$$\tau^1 = \tau \oplus \tau^2. \tag{41}$$

If $w > 0$, and $L(n) < \infty$ for all n, then Ψ^1 is the output from a $(g \star \mathbf{I}_w)^*$-server for Ψ.

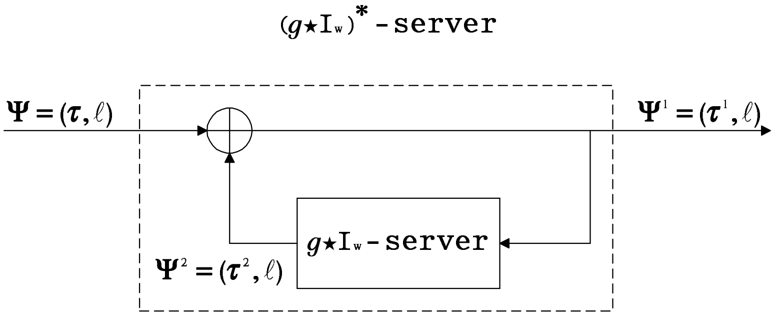

Figure 8: Feedback of a server

Proof As Ψ^2 is the output from a $g \star \mathbf{I}_w$-server for Ψ^1,

$$\tau^2 \leq \tau^1 \star_L (g \star \mathbf{I}_w).$$

From the monotonicity of $\oplus$, we have from (41) that

$$\tau^1 \leq \tau \oplus (\tau^1 \star_L (g \star \mathbf{I}_w)). \tag{42}$$

Iterating (42) yields

$$\begin{aligned}
\tau^1 &\leq \tau \oplus \Big((\tau \oplus (\tau^1 \star_L (g \star \mathbf{I}_w))) \star_L (g \star \mathbf{I}_w) \Big) \\
&\leq \tau \oplus (\tau \star_L (g \star \mathbf{I}_w)) \oplus \Big((\tau^1 \star_L (g \star \mathbf{I}_w)) \star_L (g \star \mathbf{I}_w) \Big) \\
&\leq \tau \oplus (\tau \star_L (g \star \mathbf{I}_w)) \oplus (\tau^1 \star_L (g \star \mathbf{I}_w)^{(2)}),
\end{aligned}$$

where we use the distributivity and Lemma 2.1. Thus, iterating (42) for m times yields

$$\tau^1 \leq \tau \oplus (\tau \star_L (g \star \mathbf{I}_w)) \oplus \ldots \oplus (\tau \star_L (g \star \mathbf{I}_w)^{(m)}) \oplus (\tau^1 \star_L (g \star \mathbf{I}_w)^{(m+1)}).$$

As the operator $\star$ is commutative, we have from (6) that

$$(g \star \mathbf{I}_w)^{(m+1)} = g^{(m+1)} \star \mathbf{I}_{(m+1)w}. \tag{43}$$

Since we assume that $L(n) < \infty$ and $w > 0$, for any finite n there is an m large enough such that $L(n) < (m+1)w$. Thus, we have from (7) that for all $k \leq L(n) < (m+1)w$

$$(g^{(m+1)} \star \mathbf{I}_{(m+1)w})(k) = -\infty.$$

This in turn implies that

$$(\tau^1 \star_L (g \star \mathbf{I}_w)^{(m+1)})(n) = (\tau^1 \star_L (g^{(m+1)} \star \mathbf{I}_{(m+1)w}))(n) = -\infty$$

for all n that satisfies $L(n) < (m+1)w$. For such n, we have from (43) that

$$
\begin{aligned}
&\tau^1(n) \le \Big(\tau \oplus (\tau \star_L (g \star \mathbf{I}_w)) \oplus \ldots \oplus (\tau \star_L (g \star \mathbf{I}_w)^{(m)})\Big)(n) \\
&= \Big((\tau \star_L \mathbf{I}_0) \oplus (\tau \star_L (g \star \mathbf{I}_w)) \oplus \ldots \oplus (\tau \star_L (g \star \mathbf{I}_w)^{(m)})\Big)(n) \\
&= \Big(\tau \star_L (\mathbf{I}_0 \oplus (g \star \mathbf{I}_w) \oplus \ldots \oplus (g \star \mathbf{I}_w)^{(m)}\Big)(n) \\
&\le (\tau \star_L (g \star \mathbf{I}_w)^*)(n),
\end{aligned}
$$

where in the last inequality we use the monotonicity of $\star_L$ and the definition of the closure in (8). This shows that the feedback system is indeed a $(g \star \mathbf{I}_w)^*$-server. ■

Example 4.15 (Window flow control) Let $\Psi = (\tau, \ell)$ and $\Psi^1 = (\tau^1, \ell)$ be the input and the output to a network. Assume that the service requirements are all fixed 1, i.e., $\ell(n) = 1$ for all n. This in turn implies $L(n) = n$ for all n. Let $q(t)$ be the total number of customers in the network at time t. A window flow control mechanism with window size w works as follows: a customer after it arrives is allowed to enter the network at time t if $q(t) \le w$ immediately after its entrance. Thus, in a window flow controlled network with window size w, the total number of customers in the network never exceeds w. To avoid triviality, we assume that $w > 0$. To model such a mechanism, let $\Psi^2 = (\tau^2, \ell)$ with

$$\tau^2 = \tau^1 \star \mathbf{I}_w. \tag{44}$$

The effective input to the network $\Psi^3 = (\tau^3, \ell)$ is then controlled by

$$\tau^3 = \tau \oplus \tau^2. \tag{45}$$

Note from (44) that $\tau^2(n) = -\infty$ if $n < w$, and $\tau^2(n) = \tau^1(n-w)$ if $n \ge w$. Thus, the $n+1^{th}$ customer is allowed to enter the network at

$$\tau^3(n) = \begin{cases} \tau(n) & \text{if } n < w \\ \max[\tau(n), \tau^1(n-w)] & \text{if } n \ge w \end{cases}.$$

Suppose that the network is a g-server for the effective input Ψ^3, i.e.,

$$\tau^1 \le \tau^3 \star g. \tag{46}$$

In view of (44)-(46), we have

$$\begin{aligned}\tau^3 &= \tau \oplus \tau^2 = \tau \oplus (\tau^1 \star \mathbf{I}_w) \\ &\le \tau \oplus ((\tau^3 \star g) \star \mathbf{I}_w) \\ &= \tau \oplus (\tau^3 \star (g \star \mathbf{I}_w))\end{aligned}$$

where we apply Lemma 2.1 in the last equality. From Theorem 4.14, it follows that Ψ^3 is the output from a $(g \star \mathbf{I}_w)^*$-server for Ψ. Since Ψ^1 is the output from a g-server for Ψ^3, applying Theorem 4.5 yields that Ψ^1 is the output from a $(g \star \mathbf{I}_w)^* \star g$-server for Ψ. Thus, the window flow control system is a $(g \star \mathbf{I}_w)^* \star g$-server.

5 Conclusions and future research

In this paper, we developed a general framework for providing deterministic service guarantees in telecommunication networks with variable length packets. In this framework, we proposed a traffic characterization, called g-regularity, for marked point processes. Analogous to the filtering theory under the $(\min, +)$-algebra in [5, 1, 14, 21], we introduced two basic network elements under the $(\max, +)$-algebra: (i) traffic regulators that generate g-regular marked point processes, and (ii) g-servers that provide quality-of-service for marked point processes. We also developed the rules for network elements to be joined by concatenation, "filter bank summation," and feedback to form a composite network element. The usage of these rules were illustrated by various examples that include $G/G/1$ queues, guaranteed rate servers in tandem [20], segmentation and reassembly, jitter control, dampers [12], and window flow control [1, 14, 5].

In the following, we discuss possible extensions of the framework:

(i) Joint traffic regulation and service guarantees: as the matrix extension for the filtering theory under the $(\min, +)$-algebra in [6], one may consider networks with multiple inputs. For instance, for K marked pointed processes $\Psi_k = (\tau_k, \ell_k)$, $k = 1, \ldots, K$, one may define the following joint traffic characterization:

$$\tau_i(n) - \tau_j(m) \ge G_{ij}(L_i(n) - L_j(m)), \quad i, j = 1, \ldots, K,\ m \le n, \tag{47}$$

for some $G_{ij} \in \mathcal{F}_{inc}$. Equivalently,

$$\tau_i = (\tau_1 \star_{L_1} G_{i1}) \oplus \ldots \oplus (\tau_K \star_{L_K} G_{iK}),\ i = 1, \ldots, K.$$

Similarly, one may define joint service guarantees by reversing the inequalities in (47).

(ii) Stochastic extensions: for stochastic extensions, one may use the exponential transformation in [3, 4] that transforms the (max, +)-algebra into the usual algebra. This might lead to exponential bounds under the Gärtner-Ellis large deviation conditions [18, 16] for arrival times and service requirements. For Markov processes, it is expected that one might only need g-servers with g being linear. However, for long range dependent inputs (see e.g., [22]), additional service efforts are required and one might need nonlinear g-servers as discussed in [15, 7, 8].

(iii) Dynamic service guarantees: as discussed in Section 3.1, our traffic characterization is shift invariant. This can be relaxed by considering the following dynamic traffic characterization [11]

$$\tau(n) - \tau(m) \geq G(L(n), L(m)),$$

where $G(\cdot, \cdot)$ is a function with two variables and $G(s, t)$ is increasing in t. Dynamic service guarantees can be defined similarly.

References

[1] R. Agrawal and R. Rajan, "Performance bounds for guaranteed and adaptive services," *IBM RC 20649*, 1996. Also in *Proc. 34th Allerton Conf. on Comm., Cont. & Comp.*, Monticello, IL, Oct. 1996.

[2] F. Baccelli, G. Cohen, G. Oslder, and J. Quadrat, *Synchronization and Linearity: An Algebra for Discrete Event Systems,* Wiley, 1992.

[3] C.S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, Vol.39, pp. 913-931, 1994.

[4] C.S. Chang, "On the exponentiality of stochastic linear systems under the max-plus algebra," *IEEE Transactions on Automatic Control*, Vol. 41, pp. 1182-1188, 1996.

[5] C.S. Chang, "On deterministic traffic regulation and service guarantees: a systematic approach by filtering," *IEEE Transactions on Information Theory*, Vol. 44, pp. 1097-1110, 1998. Abbreviated version "A filtering theory for deterministic traffic regulation" in *IEEE INFOCOM'97*, Kobe, Japan.

[6] C.S. Chang, "Matrix extensions of the filtering theory for deterministic traffic regulation and service guarantees," *IEEE Journal on Selected Areas of Communications*, Vol. 16, pp. 708-178, 1998.

[7] C.S. Chang, D. Yao and T. Zajic, "Large deviations, moderate deviations, and queues with long-range dependent input," *Adv. Appl. Prob.*, Vol. 31, pp. 254-278, 1999.

[8] C.S. Chang, D. Yao and T. Zajic, "Moderate deviations for queues with long-range dependent input." In: *Stochastic Networks: Stability and Rare Events*, P. Glasserman, K. Sigman and D.D. Yao (eds.), Chapter 14, Springer-Verlag, New York, 1996.

[9] R.L. Cruz, "A calculus for network delay, Part I: Network elements in isolation," *IEEE Transactions on Information Theory*, Vol. 37, pp. 114-131, 1991.

[10] R.L. Cruz, "A calculus for network delay, Part II: Network analysis," *IEEE Transactions on Information Theory*, Vol. 37, pp. 132-141, 1991.

[11] R.L. Cruz, *personal communication*, 1997.

[12] R.L. Cruz, " SCED+: efficient management of quality of service guarantees," *IEEE INFOCOM'98.*

[13] R.L. Cruz, *Lecture Notes on Quality of Service Guarantees*, 1998.

[14] R.L. Cruz and C.M. Okino, "Service guarantees for window flow control," *Proc. 34th Allerton Conf. on Comm., Cont. & Comp.*, Monticello, IL, Oct. 1996.

[15] N.G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single-server queue, with applications," *Math. Proc. Cambridge Philos. Soc*, Vol. 118, No. 2, pp. 363-374, 1995.

[16] R. Ellis, "Large deviations for a general class of random vectors, " *Ann. Probab.*, Vol. 12, pp. 1-12, 1984.

[17] N.R. Figueria and J. Pasquale, "An upper bound on delay for virtualclock service displine," *IEEE/ACM Transactions on Networking*, Vol. 3, pp. 399-408, 1994.

[18] J. Gärtner, "On large deviations from invariant measure," *Theory Probab. Appl.*, Vol. 22, pp. 24-39, 1977.

[19] L. Georgiadis, R. Guerin, and V. Peris, "Efficient network QoS provisioning based on per node traffic shaping," *IEEE/ACM Transactions on Networking*, Vol. 4, pp. 482-501, 1996.

[20] P. Goyal and H.M. Vin, "Generalized guaranteed rate scheduling algortihms: a framework," *IEEE/ACM Transactions on Networking*, Vol. 5, pp. 561-571, 1997.

[21] J.Y. Le Boudec, " Application of network calculus to guaranteed service networks," *IEEE Transactions on Information Theory*, Vol. 44, pp. 1087-1096, 1998.

[22] W.E. Leland, M.S. Taqqu, W. Willinger, D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version),"*IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, Feb. 1994.

[23] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications.* New York: Academic Press, 1979.

[24] A.K. Parekh and R.G. Gallager, "A generalized processor sharing approach to flow control in integrated service networks: the single-node case," *IEEE/ACM Transactions on Networking*, Vol. 1, pp. 344-357, 1993.

[25] H. Sariowan, R.L. Cruz and G.C. Polyzos, "SCED: a generalized scheduling policy for guaranteeing quality-of-service," submitted to *IEEE/ACM Transactions on Networking*.

[26] G.G. Xie and S.S. Lam, "Delay guarantee of virtual clock server," *IEEE/ACM Transactions on Networking*, Vol. 3, pp. 683-689, 1995.

[27] L. Zhang, "Virtualclock: a new traffic control algorithm for packet switching networks," *ACM Transactions on Computer Systems*, Vol. 9, pp. 101-124, 1991.