Stability, Queue Length and Delay of Deterministic and Stochastic Queueing Networks *

Cheng-Shang Chang
IBM Research Division
T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598
cschang@watson.ibm.com

Feb. 1992; revision Feb. 1993

Abstract

Motivated by recent development in high speed networks, in this paper we study two types of stability problems: (i) conditions for queueing networks that render bounded queue lengths and bounded delay for customers, and (ii) conditions for queueing networks in which the queue length distribution of a queue has an exponential tail with rate θ. To answer these two types of stability problems, we introduce two new notions of traffic characterization: minimum envelope rate (MER) and minimum envelope rate with respect to θ. Based on these two new notions of traffic characterization, we develop a set of rules for network operations such as superposition, input-output relation of a single queue, and routing. Specifically, we show that (i) the MER of a superposition process is less than or equal to the sum of the MER of each process, (ii) a queue is stable in the sense of bounded queue length if the MER of the input traffic is smaller than the capacity, (iii) the MER of a departure process from a stable queue is less than or equal to that of the input process (iv) the MER of a routed process from a departure process is less than or equal to the MER of the departure process multiplied by the MER of the routing process. Similar results hold for MER with respect to θ under a further assumption of independence. These rules provide a natural way to analyze feedforward networks with multiple classes of customers. For single class networks with nonfeedforward routing, we provide a new method to show that similar stability results hold for such networks under the FCFS policy. Moreover, when restricting to the family of two-state Markov modulated arrival processes, the notion of MER with respect to θ is shown

*IEEE Trans. Automatic Control, Vol. 39, pp. 913-931, 1994.

to be equivalent to the recently developed notion of effective bandwidth in communication networks.
Keywords: stability, queueing networks, effective bandwidth, large deviation

1 Introduction

As information technology advances, the demands for new types of communication services have been rapidly increased. To cope with these new demands, recent development of communication networks aims to serve these different demands through an integrated network, i.e., an Integrated Services Digital Network (ISDN). Before entering an ISDN, each service proposes to the network controller a service request which includes the information of source, destination, traffic pattern and grade of service (GOS). Judging from the "state" (current utilizations) of the network, the network controller then grants this service request if the GOS of this request and other traffic that are currently being served are satisfied. Otherwise, the service request is rejected. As noted in [42, 40, 34], an open and challenging problem is how to design a network controller to make such a decision. Recent research in this area can be found in [5, 14, 15, 19, 21, 22, 20, 23, 25, 26, 30] and many others.

Motivated by the problem in communication networks, in this paper we study two types of stability problems: (i) conditions for queueing networks that render bounded queue lengths and bounded delay for customers, and (ii) conditions for queueing networks in which the queue length distribution of a queue has an exponential tail with rate θ. The first type of problem corresponds to the case when GOS requires no loss or bounded delay, while the second type of problem might be suitable for the case when GOS requires an extremely small loss probability, e.g. 10^{-9}. Other interesting applications of these stability problems include investigation of inventory levels and due dates for manufacturing systems, especially for the semiconductor manufacturing systems [12, 33, 32]. In such systems, the variability of arrivals of parts and processing times of parts are small. Moreover, there are due dates for certain types of products. The question is if there exists a scheduling policy that meets all the due dates while keeping the inventory levels bounded.

We will answer these two types of stability problems in Sections 2 and 3 respectively. Our approach to these two problems is to develop stability conditions parallel to the classical conditions for queues with random inputs. It is well known (e.g., [31, 6, 3, 7]) that a single-server queue with interarrival times $\{T_n, n \geq 1\}$ and service times $\{S_n, n \geq 1\}$ is stable, i.e., the delays converge in distribution to a *finite* random variable, if $\{(T_n, S_n), n \geq 1\}$ is stationary and ergodic and $ET_n > ES_n$ (a random variable is finite if $\text{Prob}(|X| < \infty) = 1$). Moreover, if $ET_n < ES_n$, then the delays converge almost surely to infinity. From this classical example, we deduce two conditions: (i) traffic characterization and (ii) traffic condition. Stationarity and ergodicity reduces the complexity of characterizing a process to a single number, the average rate. The traffic condition then ensures that the input rate is smaller than the output rate.

To obtain an appropriate traffic characterization of the first type of problem, we use the notion of "envelope process" by Cruz [14, 15]. The notion of envelope process is similar to stationarity since it bounds the original process for an arbitrary shift of time (note that a stochastic process is *stationary* if its joint distribution is invariant with respect to an arbitrary shift of time). By adding the subadditive property to an envelope process, we show that a subadditive envelope process has an average rate. The subadditive property corresponds to the ergodic property in a $G/G/1$ queue, which requires the existence of an identical average for each sample path. Among all envelope processes, we denote the smallest envelope process as the minimum envelope process (MEP). The MEP is subadditive and has an average rate, denoted as the minimum envelope rate (MER). Based on the new traffic characterization of MER, we develop a set of rules for network operations such as superposition, input-output relation of a single queue, and routing.

(i) (Lemma 2.3) The MER of a superposition process is less than or equal to the sum of the MER of each process.

(ii) (Theorem 2.4) A queue is stable in the sense of bounded queue length and bounded delay for customers if the MER of the input traffic is smaller than the capacity, and it cannot be stable if the MER is larger than the capacity.

(iii) (Lemma 2.6) The MER of a departure process from a stable queue is less than or equal to that of the input process.

(iv) (Theorem 2.7) The MER of a routed process from a departure process is less than or equal to the MER of the departure process multiplied by the MER of the routing process.

These rules are parallel to classical stability results and provide a natural way to analyze feedforward networks with multiple classes of customers. The analogy is shown in Table 1.

Table 1. Analogy of stability conditions

	shift invariant	existence of average	stable	unstable
classical	stationary process	ergodicity	$ET_n > ES_n$	$ET_n < ES_n$
deterministic	envelope process	subadditivity	MER $< c$	MER $> c$

For single class networks with nonfeedforward routing, we use the facts that the MER's of departure processes are bounded by capacities in such a network and that the total number of customers in the network is decreasing with respect to the capacity of each queue. We show that the queue length and delay at each queue is bounded under the First Come First Served (FCFS) policy if the input rates from solving traffic equations are smaller than the capacities. As in Lu and Kumar [32], in general the same stability result may not hold for *multiclass* networks with nonfeedforward routing, even though the input rate is smaller than the capacity at each queue. We then discuss various scheduling policies that stabilize multiclass networks with nonfeedforward routing such as priority assignments and capacity partitions. Based on the argument for single class networks, we provide a sufficient condition for the stability of a multiclass network under the FCFS policy.

In Section 3, we generalize the notion of MER as a function of θ. This characterization is called minimum envelope rate with respect to θ. This rate function is increasing in θ and ranges between average rate and peak rate. Moreover, when restricting to the family of two-state Markov modulated arrival processes, the MER with respect to θ is shown to be equivalent to the recently developed notion of effective bandwidth in communication networks. Parallel to the development for the first type of problem, we derive a set of rules for network operations.

(i) (Lemma 3.4) The MER with respect to θ of a superposition of *independent* processes is less than (resp. equal to) the sum of the MER with respect to θ of each process (resp. when a set of large deviation conditions, $[C1-3]$ in Section 3.1, are satisfied).

(ii) (Theorems 3.8 and 3.9) If the MER with respect to θ of the input traffic is smaller than the capacity, then the queue length distribution has an exponential tail with rate θ. Moreover, the MER with respect to θ of the departure process is less than or equal to that of the input process.

(iii (Lemma 3.11) The MER with respect to θ of a routed process from a departure process can be bounded by a function of the MER with respect to θ of the departure process and the MER with respect to θ of the routing process.

These rules allow us to analyze acyclic networks with multiple classes of customers, where the arrival processes in front of each queue are independent. For a single class nonfeedforward network, we show similar result holds when the routing sequences are i.i.d. Bernoulli random variables.

We conclude the paper in Section 4, where we discuss possible extensions of the theory developed in this paper.

Throughout we use increasing and convex in the nonstrict sense.

2 Deterministic networks

In this section, we will answer the type of stability problem regarding bounded queue lengths and bounded delay for customers. We will introduce the notions of envelope processes and envelope rate in Section 2.1 as a method for traffic characterization. Network operation rules for this characterization are developed for a single queue in Section 2.2, and for a feedforward network with multiple classes of customers in Section 2.3. A single class nonfeedforward network is addressed in Section 2.4.

2.1 Envelope processes and envelope rates

Consider a nonnegative sequence $\{a(t), t = 0, 1, 2, \ldots\}$. Let $A(t_1, t_2) = \sum_{t=t_1}^{t_2-1} a(t)$. Cruz [14] introduced the following characterization of the burstiness of the sequence $a(t)$. He considered a bounding process $\hat{A}(t)$ with the following property:

$$A(t_1, t_2) \leq \hat{A}(t_2 - t_1), \ \ \forall t_1 \leq t_2.$$

This process $\hat{A}(t)$ will be called an *envelope process* of $a(t)$ in this paper. Note that $\hat{A}(t)$ is "stationary" in the sense that it only depends on the difference of the two time epochs t_1 and t_2. In the following lemma, we establish monotonicity and subadditivity for envelope processes. Recall that a process $\hat{A}(t)$ is subadditive if $\hat{A}(t_1 + t_2) \leq \hat{A}(t_1) + \hat{A}(t_2)$ for all t_1 and t_2.

Lemma 2.1 *Given that $\hat{A}(t)$ is an envelope process of some unknown nonnegative process $a(t)$, one could obtain from $\hat{A}(t)$ another envelope process $\hat{A}'(t)$ that is increasing and subadditive.*

Proof. Let $\hat{A}''(t) = \inf_{s \geq t} \hat{A}(s)$. Clearly, $\hat{A}''(t)$ is increasing. Since $a(t)$ is nonnegative, $A(t_1, t_2) \leq A(t_1, t_2 + s) \leq \hat{A}(s + t_2 - t_1)$ for all $s \geq 0$. Thus, $A(t_1, t_2) \leq \inf_{s \geq 0} \hat{A}(s + t_2 - t_1) = \hat{A}''(t_2 - t_1)$ for all $t_1 \leq t_2$ and $\hat{A}''(t)$ is an envelope process of $a(t)$.

To show the subadditivity, we construct $\hat{A}'(t)$ from $\hat{A}''(t)$ recursively by the following equation:

$$\hat{A}'(t) = \min \left[\hat{A}''(t), \min_{0<s<t} [\hat{A}'(s) + \hat{A}'(t - s)]\right].$$

It is easy to verify inductively that $\hat{A}'(t)$ is still increasing in t. Note that

$$\begin{aligned} \hat{A}'(t_1 + t_2) &\leq \min_{0<s<t_1+t_2} \left[\hat{A}'(s) + \hat{A}'(t_1 + t_2 - s)\right] \\ &\leq \hat{A}'(t_1) + \hat{A}'(t_2). \end{aligned}$$

Thus, $\hat{A}'(t)$ is subadditive. Now we show that $\hat{A}'(t)$ is an envelope process by induction on t. Clearly, it holds for $t = 1$ since $\hat{A}'(1) = \hat{A}''(1)$. Assume it holds for $t - 1$ as our induction hypothesis. From the induction hypothesis, it follows that that for all τ and $0 < s < t$,

$$A(\tau, \tau + t) = A(\tau, \tau + s) + A(\tau + s, \tau + t) \leq \hat{A}'(s) + \hat{A}'(t - s).$$

This implies that $A(\tau, \tau + t) \leq \min_{0<s<t} [\hat{A}'(s) + \hat{A}'(t - s)]$. In conjunction with $A(\tau, \tau + t) \leq \hat{A}''(t)$, we have $A(\tau, \tau + t) \leq \hat{A}'(t)$. This completes the argument for t. □

According to Lemma 2.1, we may assume that $\hat{A}(t)$ is increasing and subadditive. It is known (see [28]) that

$$\lim_{t\to\infty} \frac{\hat{A}(t)}{t} = \inf_{t\geq 1} \frac{\hat{A}(t)}{t} \stackrel{\text{def}}{=} \hat{a}$$

if $\hat{A}(t)$ is subadditive. The limit $\hat{a}$ will be referred as the *envelope rate* of the envelope process $\hat{A}(t)$.

Since envelope processes are not unique, it is natural to ask if there is a minimum one, i.e., an envelope process $A^*(t)$ satisfying $A^*(t) \leq \hat{A}(t)$ for all t and for all envelope processes $\hat{A}(t)$. Clearly, the answer to this question is

$$A^*(t) = \sup_{s\geq 0} A(s, s+t). \tag{1}$$

Hereafter, we refer to the process $A^*(t)$ as the minimum envelope process (MEP) of $a(t)$. It is easy to see that $A^*(t)$ is increasing and subadditive. Define the *minimum envelope rate* (MER) a^* as the limit, $\lim_{t\to\infty} \frac{A^*(t)}{t} = \inf_{t\geq 1} \frac{A^*(t)}{t}$. One can also view the MER by considering the family of linear envelope processes proposed by Cruz [14, 15]

$$\mathcal{F} \stackrel{\text{def}}{=} \{\hat{a} : A^*(t) \leq \hat{a}t + \hat{\sigma} \text{ for some nonnegative constant } \hat{\sigma}\}. \tag{2}$$

The linear envelope processes in (2) have been used in [14, 15] as a tool for computing the bound for delays. Clearly, $a^* \leq \hat{a}$ for all $\hat{a} \in \mathcal{F}$. Since $\lim_{t\to\infty} A^*(t)/t = a^*$, for every $\epsilon > 0$ there exists a constant t_0 such that for all $t \geq t_0$, $A^*(t)/t \leq (a^* + \epsilon)$. Let $\sigma = \max_{t<t_0}[A^*(t)] = A^*(t_0 - 1)$. It then follows that $A^*(t) \leq (a^* + \epsilon)t + \sigma$. Thus,

$$a^* = \inf[\hat{a} : \hat{a} \in \mathcal{F}]. \tag{3}$$

If the average rate of $a(t)$ exists, i.e., $\lim_{t\to\infty} \frac{A(s,s+t)}{t} = a'$ for all s, then one might ask if $a' = a^*$. In the following, we show by a counterexample that this is in general not true.

Example 2.2 Let $a(t)$ be a function that alternates between ones and zeros as follows: 1 one, 1 zero, 2 ones, 2 zeros, 3 ones, 3 zeros, 4 ones, 4 zeros, etc. Then we have

$$\frac{1}{2}t \leq A[0,t] \leq \frac{1}{2}t + \sqrt{t} + 1.$$

Thus, $\lim_{t\to\infty} \frac{A(s,s+t)}{t} = 1/2$. However, one could find a subsequence with an arbitrary number of consecutive 1's and thus $a^* = 1$.

We note that under the uniformly convergent condition, one could interchange the limit with the supremum to derive that $a^* = a'$. One could also verify that the uniformly convergent condition is satisfied when $a(t)$ is periodic.

In the following lemma, we establish bounds for the MEP and the MER of a superposition of K processes.

Lemma 2.3 *Let $a(t) = \sum_{k=1}^{K} a_k(t)$ be a superposition of K nonnegative processes. Then $A^*(t) \le \sum_{k=1}^{K} A_k^*(t)$ and $a^* \le \sum_{k=1}^{K} a_k^*$.*

Proof. Observe that

$$A^*(t) = \sup_{s\geq 0} A(s, s+t) = \sup_{s\geq 0} \sum_{k=1}^{K} A_k(s, s+t) \le \sum_{k=1}^{K} \sup_{s\geq 0} A_k(s, s+t) = \sum_{k=1}^{K} A_k^*(t). \tag{4}$$

That $a^* \le \sum_{k=1}^{K} a_k^*$ follows immediately by taking limits. $\square$

Throughout, we shall use a lower case letter to denote a process, e.g., $a(t)$ and the corresponding upper case letter to denote the partial sums, e.g., $A(t_1, t_2) = \sum_{t=t_1}^{t_2-1} a(t)$. A superscript $*$ on the corresponding upper (lower) case letter will denote the MEP (MER) of that process, e.g., $A^*(t)$ (a^*). We shall use the letters a, b and p to denote an arrival process, a departure process and a sequence of routing parameters, respectively.

2.2 A single queue

In this section, we consider a discrete-time queue with one class of customers. Let $a(t)$ and $q(t)$ be the number of arrivals at time t and the number of customers in the queue at time t respectively. Assume that the buffer size is infinite and that the server can serve c customers per unit of time. The constant c will be referred to as the capacity of the server. Under a work-conserving policy, i.e., a policy that does not allow idling when there are customers in the queue, the queue is governed by the following Lindley's equation:

$$q(t+1) = (q(t) + a(t) - c)^+ \tag{5}$$

where $(x)^+ \stackrel{\text{def}}{=} \max(0, x)$.

Let $A(t_1, t_2) = \sum_{t=t_1}^{t_2-1} a(t)$ be the number of arrivals in $[t_1, t_2)$ and $A^*(t)$ be its MEP with MER a^*. Note that $A^*(t)$ is the maximum number of arrivals within t units of time.

In the following theorem, we show that there exists a bounded delay if a^* is less than the capacity and the delay cannot be bounded if a^* exceeds the capacity. A similar result also holds for queue length.

Theorem 2.4 **(i)** *If $a^* < c$, then there exists a constant $d < \infty$ such that the delay of every customer is not longer than d.*

(ii) *If $a^* > c$, then there does* not *exist a constant $d < \infty$ such that the delay of every customer is not longer than d.*

As noted in the introduction, we have complete analogy to the the classical stability conditions for queues with random input: (i) envelope processes, which bounds the number of arrivals with respect to an arbitrary shift of time, correspond to stationary processes which require the joint distributions to be invariant with respect to an arbitrary shift of time, (ii) subadditivity of envelope processes, which guarantees the existence of a limit, corresponds to ergodicity of stationary processes which also guarantees the existence of an identical limit for every sample path, and (iii) the condition $a^* < c$ in Theorem 2.4 is simply the usual traffic condition.

Proof. (i) We first show that the length of each busy period is bounded above by a constant d. This in turn implies that the delay of each customer is bounded above by d. This argument has been used in Cruz [14, 15] and Kurose [30]. Let

$$d \stackrel{\text{def}}{=} \inf\{t \geq 1 : A^*(t) - ct \leq 0\}. \tag{6}$$

Since $\lim_{t\to\infty} \frac{A^*(t)}{t} = a^* < c$, $\lim_{t\to\infty} A^*(t) - ct = -\infty$ and thus d is finite. Observe that the total number of arrivals within d units of time is bounded above by $A^*(d)$. Thus, if we start from an empty system at time 0, then the next time (under a work conserving policy) that the queue becomes empty must be within d units of time. Following the same argument shows that each busy period is bounded above by d.

(ii) We show that the queue length cannot be bounded above by a constant. Since the server can serve at most c customers per unit of time, this in turn implies that the delay of each customer cannot be bounded above by a constant. Since $\lim_{t\to\infty} \frac{A^*(t)}{t} = \inf_{t\geq 1} \frac{A^*(t)}{t} = a^*$, we have that $A^*(t)/t \geq a^*$ for all t. From (1), it follows that for every t and $\epsilon > 0$, there exists a constant $m \geq t$ such that

$$A(m-t, m) \geq A^*(t) - \epsilon \geq a^* t - \epsilon. \tag{7}$$

If the queue is empty at time 0, expanding (5) recursively yields

$$q(m) = \max\Big[0, a(m-1)-c, a(m-1)+a(m-2)-2c, \ldots, a(m-1)+a(m-2)+\ldots a(0)-mc\Big]. \tag{8}$$

In conjunction with (7), we conclude that

$$\begin{aligned} q(m) &\geq A(m-t,m) - tc \\ &\geq (a^* - c)t - \epsilon. \end{aligned}$$

Since $a^* > c$ and t and ϵ are arbitrary, the queue length cannot be bounded above by a constant. $\square$

Remark 2.5 Though Theorem 2.4(i) is only stated for a queue with a fixed capacity c, it can be extended to a queue with a time varying capacity. Specifically, let $c(t)$ denote the maximum number of customers that can be served at time t and $C(t_1, t_2) = \sum_{t=t_1}^{t_2-1} c(t)$. Instead of defining an envelope process from above, one can also define an envelope process from below. The maximum lower envelope process, denoted by $C_*(t)$, is then defined to be $\inf_{s\geq 0} C(s, s+t)$. Analogous to the argument for MEP, one can easily verify that the maximum lower envelope process is increasing and superadditive. Thus, one can define the maximum lower envelope rate, denoted by c_*, as $\lim_{t\to\infty} C_*(t)/t = \sup_{t\geq 1} C_*(t)/t$. Under the condition $a^* < c_*$, the delay of every customer is bounded above by d, where

$$d \stackrel{\text{def}}{=} \inf\{t \geq 1 : A^*(t) - C_*(t) \leq 0\}. \tag{9}$$

In particular, if $c(t)$ is a periodic sequence, then $c^* = c_*$ and both rates are the same as its average rate. We will not pursue the notion of lower envelope processes any further in this paper. Further development along this line can be found in [10, 11].

In Theorem 2.4(i), we do not assume any particular scheduling policy as long as it is work-conserving. If we assume that the scheduling policy is FCFS, then the bound for the delay in (6) could be tightened by considering the maximum queue length. The delay of a customer that arrives at time $t+1$ is bounded above by $\lceil (q(t+1) + a(t+1))/c \rceil$. From (8), it follows that

$$q(t+1) + a(t+1) \leq \max \Big[A^*(1), A^*(2) - c, A^*(3) - 2c, \ldots, A^*(t+2) - (t+1)c \Big]. \tag{10}$$

By the definition of d in (6), $A^*(d) \leq cd$. Thus, for $t \geq d$, we have from the subadditive property of $A^*(t)$ that

$$A^*(t) - (t-1)c \leq A^*(t-d) - (t-d-1)c + A^*(d) - cd \leq A^*(t-d) - (t-d-1)c.$$

In conjunction with (10), it follows that for all t

$$q(t+1) + a(t+1) \leq \max \Big[A^*(1), A^*(2) - c, A^*(3) - 2c, \ldots, A^*(d) - c(d-1) \Big]. \tag{11}$$

Note that the right hand side of (11) is independent of t and provides an upper bound for the maximum queue length.

As shown in Example 2.2, we also note that the two conditions (i) the existence of an average rate a' and (ii) $a' < c$ are not enough to guarantee bounded delays.

Now we consider the departure process from the queue. Let $B(t_1, t_2)$ denote the number of departures in $[t_1, t_2)$. Also let $B^*(t)$ and b^* be the corresponding MEP and MER.

Lemma 2.6 *If the delay of each customer is bounded above by a finite constant, then the MER of the departure process is the same as that of the input process, i.e., $b^* = a^*$.*

Proof. Since each arrival can be delayed by at most d units of time,

$$A(t_1, t_2) \le B(t_1, t_2 + d) \le A(t_1 - d, t_2 + d). \tag{12}$$

This implies that $A^*(t) \le B^*(t + d) \le A^*(t + 2d)$. Thus, $b^* = a^*$. □

We note that the first inequality in (12) was used in Cruz [14] to compute the delays in feedforward networks. Also, Lemma 2.6 does not depend on how the queue is operated. For instance, the input process might be a superposition process of multiple classes of customers. As long as the delay for each customer is bounded above by a constant, the result in Lemma 2.6 holds for each class of customers. In the next section, we will use the input-output relation in Lemma 2.6 to discuss the stability of feedforward networks.

2.3 Multiclass networks with feedforward routing

In this section, we consider a discrete-time queueing network with K classes of customers and I queues. We assume that the buffer sizes of these I queues are infinite and that the service requirements of these K class customers at all I queues are one unit of time. The capacity of queue i is c_i, i.e., at most c_i customers can be served at queue i per unit of time. We further assume that each queue is operated under a work-conserving policy. The number of class k, $k = 1, 2, \ldots, K$, customers that arrive at the system at time t is denoted by $a_{0,k}(t), t = 0, 1, 2, \ldots$. These arriving customers are then routed to the I queues according to a set of routing parameters $p_{0,i,k}(n), i = 1, 2, \ldots, I, k = 1, 2, \ldots, K, n = 0, 1, 2, \ldots$. The n^{th} class k customer is (resp. not) routed to queue i if $p_{0,i,k}(n) = 1$ (resp. 0). Similarly, the n^{th} departure of class k customers from queue i is (resp. not) routed to queue j if $p_{i,j,k}(n) = 1$ (resp. 0). We assume that $p_{i,j,k}(n) = 0$ for all $i \ge j$, i.e., the network is feedforward (see figure 1). We note that we do not assume that $\sum_{j=1}^{I} p_{i,j,k}(n) \le 1$. This allows us to model broadcasting.

Let $a_{i,k}(t)$ be the number of class k customers that arrives at queue i at time t and $a_i(t) = \sum_{k=1}^{K} a_{i,k}(t)$. Let $q_i(t)$ be the total number of customers in front of queue i at time t. Then we have the Lindley equation for queue i, $i = 1, \ldots, I$:

$$q_i(t+1) = (q_i(t) + a_i(t) - c_i)^+. \tag{13}$$

Let $A_{j,k}(t_1, t_2) = \sum_{t=t_1}^{t_2-1} a_{j,k}(t)$, $j = 0, 1, 2 \ldots, I$, be the number of class k arrivals at queue j (the system when $j = 0$) in $[t_1, t_2)$ and $A^*_{j,k}(t)$ and $a^*_{j,k}$ be its MEP and MER. Also let $A_j(t_1, t_2) = \sum_{k=1}^{K} A_{j,k}(t_1, t_2)$ be the total number of arrivals at queue j in $[t_1, t_2)$ and $A^*_j(t)$ be its MEP with the MER a^*_j, $j = 1, \ldots, I$. Let $P^*_{i,j,k}(m)$ be the MEP of the routing process $p_{i,j,k}(n)$, i.e.,

$$P^*_{i,j,k}(m) \stackrel{\mathrm{def}}{=} \sup_{s \geq 0} \sum_{n=s}^{s+m-1} p_{i,j,k}(n). \tag{14}$$

In other words, $P^*_{i,j,k}(m)$ is the maximum number of class k customers that are routed to queue j among m consecutive class k customers that depart from queue i. Similarly, let $p^*_{i,j,k}$ be the corresponding MER. Let

$$\rho_{j,k} = a^*_{0,k} p^*_{0,j,k} + \sum_{i=1}^{j-1} \rho_{i,k} p^*_{i,j,k}, \quad j = 1, \ldots, I. \tag{15}$$

In the following theorem, we will show that $\rho_{j,k}$ is an upper bound for the MER of class k customers at queue j and that the delay of each customer can be bounded above by a constant if the bound for the MER of the arrival process at each queue is less than the capacity of the corresponding queue.

Theorem 2.7 *If $\sum_{k=1}^{K} \rho_{j,k} < c_j$ for all $j = 1, \ldots, I$, then the delay of a customer through the network can be bounded above by a constant.*

Proof. We will prove this by double induction on $j = 1, \ldots I$. Our induction hypotheses are (i) the delay is bounded above by a constant and (ii) the MER of the departure process of class k customers from queue j, denoted by $b^*_{j,k}$, is not greater than $\rho_{j,k}$. First, we show the case $j = 1$. In this case, the number of class k customers arrived at the first queue in $[t_1, t_2)$ is the number of class k customers that arrives within this time interval and are routed to queue 1. Thus, we have

$$A_{1,k}(t_1, t_2) \leq P^*_{0,1,k}(A_{0,k}(t_1, t_2)) \leq P^*_{0,1,k}(A^*_{0,k}(t_2 - t_1)). \tag{16}$$

This implies that

$$A^*_{1,k}(t) \leq P^*_{0,1,k}(A^*_{0,k}(t)). \tag{17}$$

Note that

$$\lim_{t\to\infty}\frac{P^*_{0,1,k}(A^*_{0,k}(t))}{t}=\lim_{t\to\infty}\frac{P^*_{0,1,k}(A^*_{0,k}(t))}{A^*_{0,k}(t)}\frac{A^*_{0,k}(t)}{t}=p^*_{0,1,k}a^*_{0,k}.$$

Since $\rho_{1,k}=p^*_{0,1,k}a^*_{0,k}$, we have

$$a^*_{1,k}\le\rho_{1,k}. \tag{18}$$

In conjunction with Lemma 2.3 for a superposition of K processes, it follows that

$$a^*_1\le\sum_{k=1}^K a^*_{1,k}\le\sum_{k=1}^K\rho_{1,k}.$$

It then follows from Theorem 2.4 (i) and the assumption $\sum_{k=1}^K\rho_{1,k}<c_1$ that the delay for each customer at queue 1 is bounded above by a constant. This completes the argument for the induction hypotheses (i) for queue 1. That the induction hypothesis (ii) for queue 1 holds follows from Lemma 2.6 and (18).

Now suppose the induction hypotheses (i) and (ii) hold for the first $j-1$ queues. Note that the arrival process of class k customers at queue j is a superposition of the external arrivals that are routed directly to queue j, and all the class k customers that depart from queue i, $i=1,\dots j-1$, and are routed to queue j. Using Lemma 2.3, the argument in the previous paragraph for routing and the induction hypothesis (ii) yields

$$a^*_{j,k}\le a^*_{0,k}p^*_{0,j,k}+\sum_{i=1}^{j-1}b^*_{i,k}p^*_{i,j,k}\le a^*_{0,k}p^*_{0,j,k}+\sum_{i=1}^{j-1}\rho_{i,k}p^*_{i,j,k}=\rho_{j,k}.$$

Apply Lemma 2.3 once more to show that $a^*_j\le\sum_{k=1}^K\rho_{j,k}$. Again, it follows from Theorem 2.4 (i) and the assumption $\sum_{k=1}^K\rho_{j,k}<c_j$ that the delay at queue j is also bounded above by a finite constant. Finally, applying Lemma 2.6 completes the induction hypothesis (ii) for queue j. □

We note that the stability result in Theorem 2.7 can also be extended to networks in which customers have different service times. Let $s_{i,k}(n)$, $i=1,\dots I$, $k=1,\dots,K$, denote the service time of the n^{th} class k customer at *queue* i and $s^*_{i,k}$'s be the corresponding MER's.

Theorem 2.8 *If $\sum_{k=1}^K\rho_{j,k}s^*_{j,k}<c_j$ for all j, then the delay of a customer through the network can be bounded above by a constant.*

Proof. We only prove it for a queue with a single class of customer. The rest of the proof is completely parallel to the development in Theorem 2.7. Consider the workload process, $v(t)$, (virtual waiting times) that satisfies the following Lindley's equation:

$$v(t+1) = (v(t) + w(t) - c)^+, \tag{19}$$

where $w(t)$ is the total amount of work that arrives at time t. As in the proof for Theorem 2.4, one can show that the busy period at each queue is bounded above by a finite constant, if $w^* < c$. Since the total amount of work that arrives within an interval is the sum of the work of the customers that arrive within the interval, $\sum_{t_1}^{t_2-1} w(t)$ satisfies a similar inequality to that in (16). This implies $w^* < a^* s^*$. In conjunction with the assumption $a^* s^* < c$, the delay of every customer is bounded above by a constant. □

Note that $s_{i,k}(n)$ may not be the same as the service time of the n^{th} class k customer that arrives at the network. However, for the network with fixed routing for each class, i.e., $p_{i,j,k}(n) = 1$ or 0 for all n, and the FCFS policy at each queue, the n^{th} class k customer at queue i is also the same as the n^{th} class k customer that arrives at the network.

2.4 Single class networks with nonfeedforward routing

In this section, we consider a discrete-time queueing network similar to the one in §2.3 with the following two exceptions: (i) there is only one class of customer (and thus the index k will be dropped in this section), and (ii) the routing can be nonfeedforward, i.e., $p_{i,j}(n)$ may not be 0 for all $i < j$ (see figure 2). In a nonfeedforward network, an individual customer could be circled within the network for an arbitrary number of times. Thus, the delay for an individual customer cannot be bounded and we are interested in the conditions that result in bounded delay for *each queue*. To be precise, let $\nu_j = a_0^* p_{0,j}^*$, $j = 1, \ldots, I$. From §2.3, ν_j is an upper bound of the MER of the external arrivals to queue j. Similar to the definition of ρ_j in (15), let ρ_j, $j = 1, \ldots, I$ be the solution of the following traffic equation:

$$\rho_j = \nu_j + \sum_{i=1}^{I} \rho_i p_{i,j}^*. \tag{20}$$

As one might notice, ρ_j is the arrival rate to queue j (including both external and internal traffic) in the Jackson network with the external arrival rates ν_j, $j = 1, \ldots, I$ and the routing probabilities $p_{i,j}^*$, $i, j = 1, \ldots, I$. It is known that the Jackson network is stable if $\rho_j < c_j$ for all j. These conditions will be referred to as the usual traffic conditions. Our interest in this section is to answer if the delay at each queue can be bounded above by a finite constant under the usual traffic conditions.

Due to the possibility of customers being circled around, our inductive proof in the previous section cannot be applied to nonfeedforward networks. As a natural extension of induction, one might consider the fixed point iteration algorithm as in [15, 30]. First, one considers the network excludes all internal traffic. Each queue is analyzed in isolation and a bound is found for the MEP of the departure process from each queue. These bounds for the MEPs of the departure processes are then incorporated with the bounds for external traffic to analyze each queue in isolation. The procedure is iterated until the bounds for the MEPs of the departure processes converge. Unfortunately, as noted in Cruz [15], this iteration algorithm converges only if the traffic is sufficiently low and the traffic conditions needed for it to converge are stronger than the usual ones. Thus, a different approach is needed.

Before we introduce our approach, let us simplify the notations by using matrix representation. Let $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_I]$, $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_I]$, $\boldsymbol{c} = [c_1, \ldots, c_I]$ and $\boldsymbol{p}^*$ be a matrix with $p^*_{i,j}$ being the element in the i^{th} row and j^{th} column. Write (20) in matrix form as follows:

$$\boldsymbol{\rho} = \boldsymbol{\nu} + \boldsymbol{\rho}\boldsymbol{p}^*. \tag{21}$$

Equation (21) can be solved by the fixed point iteration. Since $\boldsymbol{\nu}$ and $\boldsymbol{p}^*$ are nonnegative, the sequence of vectors, $\boldsymbol{\rho}^{(n)} = \boldsymbol{\nu} + \boldsymbol{\rho}^{(n-1)}\boldsymbol{p}^*$ with $\boldsymbol{\rho}^{(0)} = [0., \ldots, 0]$, is increasing in n. If the spectral radius of the matrix $\boldsymbol{p}^*$, denoted by $\mathrm{sp}(\boldsymbol{p}^*)$, is less than 1, then the sequence $\boldsymbol{\rho}^{(n)}$ converges to $\boldsymbol{\rho}$ and ρ_j, $j = 1, \ldots, I$ are finite. Moreover, the matrix $\boldsymbol{\Gamma} - \boldsymbol{p}^*$ is invertible and thus $\boldsymbol{\rho} = \boldsymbol{\nu}(\boldsymbol{\Gamma} - \boldsymbol{p}^*)^{-1}$, where $\boldsymbol{\Gamma}$ is the identity matrix. A sufficient condition for $\mathrm{sp}(\boldsymbol{p}^*) < 1$ is that $\sum_{j=1}^{I} p^*_{i,j} < 1$ for all i ([24], Theorems 5.6.5 and 5.6.9). Hereafter, we will assume that $\mathrm{sp}(\boldsymbol{p}^*) < 1$.

Our approach for the stability problem consists of the following steps. We first consider two open polyhedral sets E_1 and E_2 (below) obtained from a strong traffic condition and the usual traffic condition. In Lemma 2.9, we show that a bounded delay at each queue can be achieved under the stronger traffic condition E_1. We then relax the traffic condition from E_1 to E_2 using the monotonicity result in Lemma 2.10.

Now consider the following two open polyhedral sets:

$$E_1 = \{\boldsymbol{c} : \boldsymbol{\nu} < \boldsymbol{c}(\boldsymbol{\Gamma} - \boldsymbol{p}^*)\} \tag{22}$$

$$E_2 = \{\boldsymbol{c} : \boldsymbol{\nu}(\boldsymbol{\Gamma} - \boldsymbol{p}^*)^{-1} < \boldsymbol{c}\} \tag{23}$$

It is easy to see that E_1 is a shifted cone (i.e., all the hyperplanes pass through $\boldsymbol{\rho}$) and that $\boldsymbol{\rho}$ is an extreme direction of E_1 (i.e., for any $\boldsymbol{c} \in E_1$ and $\lambda > 0$, $\boldsymbol{c} + \lambda\boldsymbol{\rho} \in E_1$). The open polyhedral set E_2 is simply the quadrant $\{\boldsymbol{c} : \boldsymbol{c} > \boldsymbol{\rho}\}$. Moreover, we have the following two properties between these two open polyhedral sets.

(P1) $E_1 \subset E_2$.

(P2) For every vector $\boldsymbol{c} \in E_2$, there is a vector $\boldsymbol{c}^1 \in E_1$ such that $\boldsymbol{c}^1 \leq \boldsymbol{c}$.

To show the first property, one observes that

$$(\boldsymbol{\Gamma} - \boldsymbol{p}^*)^{-1} = \boldsymbol{\Gamma} + \boldsymbol{p}^* + (\boldsymbol{p}^*)^2 + (\boldsymbol{p}^*)^3 + \dots \tag{24}$$

is a nonnegative matrix with positive diagonal elements since the matrix p^* is nonnegative. Thus, if a vector $\boldsymbol{c}$ is in E_1, then multiplying both sides of (22) by the matrix $(\boldsymbol{\Gamma} - \boldsymbol{p}^*)^{-1}$ yields

$$\boldsymbol{\nu}(\boldsymbol{\Gamma} - \boldsymbol{p}^*)^{-1} < \boldsymbol{c}(\boldsymbol{\Gamma} - \boldsymbol{p}^*)(\boldsymbol{\Gamma} - \boldsymbol{p}^*)^{-1} \tag{25}$$

since each scalar inequality in (25) is a linear combination of the scalar inequalities in (22) with at least one positive coefficient. This shows that $\boldsymbol{c} \in E_2$. For (P2), we only show the case $\boldsymbol{\nu} > 0$. The general case can be shown by a similar argument and the property of open sets. Since $\boldsymbol{\nu} > 0$, we have $\boldsymbol{\rho} > 0$. Let $\lambda \stackrel{\text{def}}{=} \min[c_1/\rho_1, \dots, c_I/\rho_I]$. Clearly, $\lambda > 1$ if $\boldsymbol{c} \in E_2$. Now let $\boldsymbol{c}^1 = \lambda\boldsymbol{\rho}$. It is easy to see that $\boldsymbol{c}^1 \leq \boldsymbol{c}$. Moreover, $\boldsymbol{c}^1 \in E_1$ since $\lambda\boldsymbol{\rho}(\boldsymbol{\Gamma} - \boldsymbol{p}^*) = \lambda\boldsymbol{\nu} > \boldsymbol{\nu}$. As an example, consider $\boldsymbol{\nu} = (1, 1)$ and the matrix

$$\boldsymbol{p}^* = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}.$$

In this example, $\boldsymbol{\rho} = (2, 2)$ and the regions of E_1 and E_2 are shown in figure 3.

Lemma 2.9 *If $\mathbf{c} \in E_1$, then the delay at each queue is bounded above by a finite constant and the queue length at each queue is also bounded.*

Proof. Note that the MER of the departure process from queue j is bounded above by the capacity c_j. Thus, the MER of the arrival process to queue j is bounded above by $\nu_j + \sum_{i=1}^I c_i p^*_{i,j}$. From Theorem 2.4 (i), it follows that customers at queue j would have a bounded delay if

$$\nu_j + \sum_{i=1}^{I} c_i p^*_{i,j} < c_j. \tag{26}$$

The condition (22) is the matrix form of (26). □

Lemma 2.10 *For two queueing networks described in this section, if the capacities are ordered, i.e., $\boldsymbol{c}^1 \leq \boldsymbol{c}^2$, then the number of departures from each queue by time t at the first system is not greater than that at the second system. As a direct consequence, the total number of customers in the first system is not less than that of the second system.*

Proof. This type of monotonicity result is well known in the literature (see Tsoucas and Walrand [39], Foss [18] and references therein). For completeness, we provide the argument in [39]. Let $r_{i,t}(n)$ be the time remaining at time t until the n^{th} customer that arrives at queue i leaves queue i. By convention, $r_{i,t}(n) = \infty$ if fewer than n customers have arrived at queue i by time t and $r_{i,t}(n) = 0$ if the n^{th} customer has departed from queue i by time t. Then it can proved by induction on t that $r^1_{i,t}(n) \geq r^2_{i,t}(n)$ for all i, n, t if $\boldsymbol{c}^1 \leq \boldsymbol{c}^2$. $\square$

Remark 2.11 We note that if there are non-integer, but rational components in $\boldsymbol{c}^1$ in Lemma 2.10, then those capacities should be interpreted as periodic sequences that alternate between their ceiling and floor values (cf. Remark 2.5). For instance, if $c = 2.5$, then $c(2t) = 2$ and $c(2t+1) = 3$ for all t, where $c(t)$ is the maximum number of customers that can be served per unit of time. One can easily verify that both Lemmas 2.9 and 2.10 hold for this periodic interpretation for a non-integer capacity. (To apply the sample path argument in Lemma 2.10, one should construct the two periodic sequences such that $c^1(t) \leq c^2(t)$ for all t.)

Analogous to the stability conditions for the Jackson networks, we have the following theorem:

Theorem 2.12 *If $\rho_j < c_j$ for all j and $sp(\boldsymbol{p}^*) < 1$, then every queue length can be bounded above by a finite constant. As a direct consequence, the delay at each queue is bounded if the service discipline is FCFS.*

Proof. For every $\boldsymbol{c}^2 \in E_2$, there exists $\boldsymbol{c}^1 \in E_1$ such that $\boldsymbol{c}^1 \leq \boldsymbol{c}^2$. From Lemma 2.9, the queue length at each queue is bounded above by a constant when the system has capacity $\boldsymbol{c}^1$. This implies that the total number of customers in the system is still bounded above by a finite constant. Applying Lemma 2.10, the total number of customers in the system with capacity $\boldsymbol{c}^2$ is then bounded above by the same constant. Thus, each queue is bounded above by the same constant. If, furthermore, the service discipline is FCFS, it then follows from the same argument as in (11) that the delay at each queue is bounded above by a finite constant. $\square$

Corollary 2.13 *If $\rho_j < c_j$, $sp(\boldsymbol{p}^*) < 1$, and the service discipline at each queue is FCFS, then $b^*_j \leq \rho_j$ for all j, where b^*_j is the MER of the departure process from queue j.*

Proof. The MER of the arrival process from queue i to queue j is bounded above by $b^*_i p^*_{i,j}$. Thus, the MER of the arrival process to queue j is bounded above by $\nu_j + \sum_{i=1}^{I} b^*_i p^*_{i,j}$. Since we assume that $\rho_j < c_j$, it then follows from Theorem 2.12 the delay at each queue is bounded above by a constant. In conjunction with Lemma 2.6, we have $b^*_j \leq \nu_j + \sum_{i=1}^{n} b^*_i p^*_{i,j}$ or equivalently

$$\boldsymbol{b}^*(\boldsymbol{\Gamma} - \boldsymbol{p}^*) \leq \boldsymbol{\nu}, \tag{27}$$

where $\boldsymbol{b}^* = [b_1^*, \ldots, b_I^*]$. Analogous to the argument for (P1), we multiply both sides of (27) by the matrix $(\boldsymbol{\Gamma} - \boldsymbol{p}^*)^{-1}$. We then have $\boldsymbol{b}^* \leq \boldsymbol{\nu}(\boldsymbol{\Gamma} - \boldsymbol{p}^*)^{-1} = \boldsymbol{\rho}$. □

We note one can also use the argument in this section to compute the bound for the total number of customers in the network. However, this bound may not be tight.

To stabilize a K-class nonfeedforward network, one can reserve a certain portion of the capacity at each queue to each class of traffic. Thus, the system behaves like K independent single class nonfeedforward networks and each one of them can be shown to be stable by the argument developed in this section. Another way to stabilize a K-class nonfeedforward network (with fixed routing) is to assign appropriate priorities to classes of jobs at each queue. For instance, one could assign priorities according to the order that queues are visited (see [32]). By so doing, the class of jobs that has higher priority is not affected by the other classes of jobs. Moreover, the traffic of this class of jobs entering its first queue is also not affected by its own internal traffic from other queues. Thus, the induction technique in §2.3 can be used to show the stability of the network under the usual traffic conditions, i.e., $\sum_{k=1}^{K} \rho_{j,k} < c_j$ for all j, where $\rho_{j,k}$ is the solution of (20) for class k customers. However, it is still not clear if the system could be stabilized under the FCFS policy when the usual traffic conditions are satisfied. The main difficulty in analyzing multiclass nonfeedforward networks is that the departure process from each queue consists of different classes of customers. If we simply bound the departure process of each class by capacity, the bound is too loose to derive the desired traffic conditions. However, we still can mimic the proof for Theorem 2.12 to obtain sufficient conditions. Recall that the number of arrivals from queue i to queue j within a time interval of t units of time is bounded above by the number of customers that depart from queue i within that interval and are routed to queue j. Suppose there are n_k class k customers that depart from queue i within the interval of t units of time. Clearly, $\sum_{k=1}^{K} n_k \leq c_i t$. Thus, we have

$$A_{i,j}^*(t) \leq \max_{n_1 + \ldots + n_K \leq c_i t} [\sum_{k=1}^{K} P_{i,j,k}^*(n_k)] \stackrel{\text{def}}{=} \hat{A}_{i,j}(t). \tag{28}$$

It is easy to verify that $\hat{A}_{i,j}(t)$ defined above is also increasing and subadditive in t. Since $\hat{A}_{i,j}(t) \geq \max_k [P_{i,j,k}^*(c_i t)]$, $\lim_{t \to \infty} \hat{A}_{i,j}(t)/t \geq c_i \max_k [p_{i,j,k}^*]$. Note that for every $\epsilon_{i,j,k} > 0$ there exists a constant $\sigma_{i,j,k}$ such that

$$P_{i,j,k}^*(n) \leq (p_{i,j,k}^* + \epsilon_{i,j,k})n + \sigma_{i,j,k}. \tag{29}$$

It then follows that $\hat{A}_{i,j}(t) \leq (\max_k [p_{i,j,k}^*] + \epsilon_{i,j}) c_i t + \sigma_{i,j}$, where $\epsilon_{i,j} = \max_k [\epsilon_{i,j,k}]$ and $\sigma_{i,j} = \max_k [\sigma_{i,j,k}]$. Thus,

$$\lim_{t \to \infty} \frac{\hat{A}_{i,j}(t)}{t} = c_i \max_k [p_{i,j,k}^*].$$

Now using the same argument as in the proof of Theorem 2.12, one can show that a multiclass nonfeedforward network under FCFS policy is stable if $\rho_j < c_j$ for all j, where ρ_j is the solution of (20) with $p^*_{i,j} = \max_k [p^*_{i,j,k}]$ and $\nu_j = \sum_{k=1}^{K} \nu_{j,k}$. However, these traffic conditions are stronger than desired.

3 Stochastic networks

In this section, we extend our results in the previous section from deterministic queueing networks to stochastic queueing networks. Our objectives in this section are (i) to provide a tool to compute simple bounds for tail distributions and (ii) to answer the second type of stability problem of queueing networks.

Instead of having deterministic bounds for random variables as in the previous section, in this section we consider bounds for moment generating functions. We say a random variable X is bounded exponentially with respect to θ ($0 < \theta < \infty$) if the θ-norm of $\exp(X)$ is finite, i.e., there exist a constant $d < \infty$ such that

$$(Ee^{\theta X})^{\frac{1}{\theta}} \leq d. \tag{30}$$

Thus, we have from Chernoff's bound that

$$P(X \geq x) \leq d^{\theta} e^{-\theta x} \quad \text{for all } x,$$

which provides a bound for the tail distribution of X.

Parallel to the development in deterministic queueing networks, we consider envelope processes (EP) of input processes with respect to θ in Section 3.1. Among the EPs, the class of linear EPs is of importance, as noted by Cruz [14, 15] in a deterministic setting. We show that if the input process in a single queue has a linear EP whose rate is smaller than the capacity, c, and the queue is operated under a work-conserving policy, then (i) the queue length is bounded exponentially with respect to θ, (ii) there exists a linear EP of the departure process which can be represented as a function of the linear EP of the input process and (iii) the virtual delay is bounded exponentially with respect to θc if the scheduling policy is First Come First Served (FCFS). Using these results, bounds for the tail distributions of queue length and virtual delay can be computed easily from the linear EP of the input process. Like in the previous section for deterministic networks, the minimum envelope rate with respect to θ (MER) is the infimum of the rates in the class of linear EPs. A sufficient condition for the queue length to be bounded exponentially with respect to θ is that the MER of the input process is smaller than the capacity. On the other hand, if the MER is larger than the capacity, then the queue length cannot be bounded exponentially with respect to θ. In particular, when the arrival process is a superposition of independent two-state Markov modulated processes, we show that the notion of MER is equivalent to the recently developed notion of effective bandwidth in [26, 20, 23] and is also related to the Perron-Frobenius eigenvalue in [37].

In order to extend these results to networks, in Section 3.3 we consider marked point processes, in which there are a sequence of arrival points and a sequence of marks associated with the arrival points. The marks can be viewed either as the service requirements or the routing variables. We show that if (i) there are a linear EP of the arrival process and a linear EP of the marking sequence and (ii) the arrival points and marks are independent, then there is a linear EP of the marked process in terms of the linear EP of the arrival process and the linear EP of the marking process. Using these input-output types of relations, we extend the bounds for the tail distributions of virtual delay and queue length from a single queue to acyclic networks, where the paths of customers do not form a loop and the input at each queue is a superposition of *independent* processes. Note that the notion of *independence*, though trivial in deterministic networks, is crucial in stochastic networks.

We then consider a *single* class network with nonfeedforward routing in Section 3.4. We assume that the routing random variables from each queue are independent and identically distributed (i.i.d.). Using an argument similar to that in Section 2.4, we show that the queue length at each queue can be bounded exponentially with respect to θ if the strong traffic condition (E_1) is satisfied. Under the weak traffic condition (E_2), we show that the total number of customers in the system can be bounded exponentially with respect to θ/I, where I is the number of queues in the network.

3.1 Envelope processes and envelope rates

Consider a sequence of *nonnegative* random variables, $\{a(t), t = 0, 1, 2, \ldots\}$. Let $A(t_1, t_2) = \sum_{t=t_1}^{t_2-1} a(t)$. Analogous to the notions of envelope processes in previous section for deterministic networks, we consider the following "bounding" process of $a(t)$:

$$\frac{1}{\theta} \log E e^{\theta A(t_1,t_2)} \leq \hat{A}(\theta, t_2 - t_1) \;\; \forall\, t_1 \leq t_2 \tag{31}$$

The process $\hat{A}(\theta, t)$ will be also called an *envelope process* of $a(t)$ with respect to θ (EP). Clearly, the minimum envelope process with respect to θ (MEP) is

$$A^*(\theta, t) = \sup_{s \geq 0} \frac{1}{\theta} \log E e^{\theta A(s,s+t)}. \tag{32}$$

Unlike the MEP in a deterministic setting, the MEP defined in (32) is not subadditive in general. Thus, we define the *minimum envelope rate* of $a(t)$ with respect to θ (MER) to be

$$a^*(\theta) = \limsup_{t\to\infty} \frac{A^*(\theta,t)}{t}. \tag{33}$$

Similar to (2), one can also view the MER by considering the family of linear EPs.

$$\mathcal{F}_\theta \stackrel{\text{def}}{=} \{\hat{a}(\theta) : A^*(\theta,t) \le \hat{a}(\theta)t + \hat{\sigma}(\theta) \text{ for some nonnegative constant } \hat{\sigma}(\theta)\}, \tag{34}$$

where $\hat{a}(\theta)$ will be called the rate of a linear EP. Note that $\hat{\sigma}(\theta)$ is constant in t, but it is a function of θ. Using the same argument as in (3), it is easy to see that for each fixed θ,

$$a^*(\theta) = \inf[\hat{a}(\theta) : \hat{a}(\theta) \in \mathcal{F}_\theta]. \tag{35}$$

We note that our definition of MER is connected to the theory of large deviation through the Gärtner-Ellis theorem. To establish the connection, we further introduce the following conditions for a sequence $\{a(t), t \ge 0\}$.

(C1) $\{a(t), t \ge 0\}$ is stationary and ergodic.

(C2) $a^*(\theta) = \lim_{t\to\infty} \frac{A^*(\theta,t)}{t}$ for all $0 < \theta < \infty$.

(C3) $\theta a^*(\theta)$ is strictly convex and differentiable for all $0 < \theta < \infty$.

Under these three conditions, the sequence $\{A(0,t), t \ge 1\}$ obeys the large deviation principle (see [8]) with the rate function

$$I(v) = \sup_\theta \{\theta v - \theta a^*(\theta)\}. \tag{36}$$

We note that $\theta a^*(\theta)$ is increasing and convex for $0 \le \theta < \infty$ according to the definition of $a^*(\theta)$. Strict convexity of $\theta a^*(\theta)$ implies that $a^*(\theta)$ is strictly increasing. We will discuss more on monotonicity and bounds in Lemma 3.5. Moreover, under $[C1-3]$ one can also verify that for all $t_1 \le t_2$ and any $\epsilon > 0$, there is a constant $\hat{\sigma}(\theta) \ge 0$ such that

$$(a^*(\theta) - \epsilon)(t_2 - t_1) - \hat{\sigma}(\theta) \le \frac{1}{\theta} \log E e^{\theta A(t_1,t_2)} \le (a^*(\theta) + \epsilon)(t_2 - t_1) + \hat{\sigma}(\theta). \tag{37}$$

Thus, $a^*(\theta)$ is not only the minimum upper envelope rate but also the maximum lower envelope rate. For further development along this line, we refer to [29, 11].

In the next section, we will first use linear EPs to derive input-output relations between arrival processes and departure processes and then apply the representation in (35) to establish stability results. Now we consider some stochastic processes where these concepts can be easily applied.

Example 3.1 If $a(t)$ is a sequence of *independent* random variables, then the MEP $A^*(\theta, t)$ is subadditive. Thus,

$$a^*(\theta) = \lim_{t\to\infty} \frac{A^*(\theta,t)}{t} = \inf_{t\geq 1} \frac{A^*(\theta,t)}{t}.$$

Proof. Observe that

$$A^*(\theta, t_1+t_2) = \sup_{s\geq 0} \frac{1}{\theta} \log Ee^{\theta A(s,s+t_1+t_2)} = \sup_{s\geq 0} \frac{1}{\theta} \log Ee^{\theta A(s,s+t_1)+\theta A(s+t_1,s+t_1+t_2)} \tag{38}$$

Thus,

$$\begin{aligned} A^*(\theta, t_1+t_2) &= \sup_{s\geq 0} \Big[\frac{1}{\theta}\log Ee^{\theta A(s,s+t_1)} + \frac{1}{\theta}\log Ee^{\theta A(s+t_1,s+t_1+t_2)}\Big] \\ &\leq \sup_{s\geq 0} \frac{1}{\theta}\log Ee^{\theta A(s,s+t_1)} + \sup_{s\geq 0}\frac{1}{\theta}\log Ee^{\theta A(s+t_1,s+t_1+t_2)} \\ &\leq A^*(\theta,t_1) + A^*(\theta,t_2). \end{aligned}$$

The limit then follows from the subadditive property (see [28]). □

In the second example, we consider stationary and associated processes. A process $a(t)$ is said to be *stationary* if its joint distribution is invariant with respect to an arbitrary shift of time, i.e.,

$$\mathrm{Prob}(a(t_1) < x_1, \ldots, a(t_n) < x_n) = \mathrm{Prob}(a(t_1+s) < x_1, \ldots, a(t_n+s) < x_n) \tag{39}$$

for all $t_1, \ldots, t_n$ and s. A process $a(t)$ is said to be associated if all the random variables, $\{a(t), t = 0, 1, 2, \ldots\}$, are associated, i.e.,

$$Ef(a(t_1),\ldots,a(t_n))g(a(t_1),\ldots,a(t_n)) \geq Ef(a(t_1),\ldots,a(t_n))Eg(a(t_1),\ldots,a(t_n)) \tag{40}$$

for all $t_1, \ldots, t_n$ and for all f, g increasing. For the properties of associated random variables, we refer to [4, 17].

Example 3.2 If $a(t)$ is stationary and associated, then the MEP $A^*(\theta, t)$ is superadditive. Thus,

$$a^*(\theta) = \lim_{t\to\infty} \frac{A^*(\theta,t)}{t} = \sup_{t\geq 1} \frac{A^*(\theta,t)}{t}.$$

Proof. Observe from stationarity that

$$A^*(\theta, t_1+t_2) = \frac{1}{\theta}\log Ee^{\theta A(0,t_1+t_2)} = \frac{1}{\theta}\log Ee^{\theta A(0,t_1)}e^{\theta A(t_1,t_2+t_2)}.$$

Since $a(t)$ is associated, the two random variables $A(0,t_1)$ and $A(t_1,t_1+t_2)$ are associated. Thus,

$$A^*(\theta,t_1+t_2) \geq \frac{1}{\theta}\log Ee^{\theta A(0,t_1)} + \frac{1}{\theta}\log Ee^{\theta A(t_1,t_1+t_2)}.$$

Since $a(t)$ is stationary, the two random variables $A(t_1,t_1+t_2)$ and $A(0,t_2)$ have the same distribution. Thus, we have

$$A^*(\theta,t_1+t_2) \geq A^*(\theta,t_1) + A^*(\theta,t_2).$$

Again, the limit then follows from the superadditive property [28]. □

As a special case of Examples 3.1 and 3.2, $a^*(\theta) = (1/\theta)\log(E\exp(\theta a(0)))$ if $a(t)$ is a sequence of i.i.d. random variables. The MER $a^*(\theta)$ for i.i.d. random variables is referred to as *effective bandwidth* in Kelly [26].

In the third example, we consider a Markov modulated process (MMP). Let $x(t)$ be a Markov process on the states $\{1,\ldots,M\}$ with the transition matrix $\boldsymbol{r}$, i.e., $r_{i,j}$ is the transition probability from state i to state j. Also let $\{y_i(t), t=0,1,\ldots\}$, $i=1,\ldots,M$, be M sequences of i.i.d. random variables with the moment generating functions $\phi_i(\theta) = E\exp(\theta y_i(0))$. The process $a(t) = y_{x(t)}(t)$ is then an MMP with the modulating process $x(t)$. Clearly, $a(t)$ is stationary if $x(t)$ is stationary.

Example 3.3 Consider an MMP $a(t)$ as described above. Let $\boldsymbol{\phi}(\theta)$ be the diagonal matrix $\mathrm{diag}\{\phi_1(\theta),\ldots,\phi_M(\theta)\}$ and $\mathrm{sp}(\boldsymbol{\phi}(\theta)\boldsymbol{r})$ be the spectral radius of the matrix $\boldsymbol{\phi}(\theta)\boldsymbol{r}$. Then the MER $a^*(\theta)$ is bounded above by $(1/\theta)\log\mathrm{sp}(\boldsymbol{\phi}(\theta)\boldsymbol{r})$.

If, furthermore, the Markov process $x(t)$ with the transition matrix $\boldsymbol{r}$ is irreducible and aperiodic, then

$$a^*(\theta) = \lim_{t\to\infty}\frac{A^*(\theta,t)}{t} = \frac{1}{\theta}\log\mathrm{sp}(\boldsymbol{\phi}(\theta)\boldsymbol{r}).$$

Note that $\mathrm{sp}(\boldsymbol{\phi}(\theta)\boldsymbol{r}) = \mathrm{sp}(\boldsymbol{r}\boldsymbol{\phi}(\theta))$ ([24], Theorem 1.3.20).

Proof. Analogous to the backward equation, one observes that

$$E(e^{\theta A(0,t)}|x(0)=i) = \phi_i(\theta)\sum_{j=1}^{M} E(e^{\theta A(0,t-1)}|x(0)=j)r_{i,j}. \tag{41}$$

Let

$$\boldsymbol{\psi}(\theta,t) = (E(e^{\theta A(0,t)}|x(0)=1),\ldots,E(e^{\theta A(0,t)}|x(0)=M))$$

and $\boldsymbol{\psi}(\theta,t)^T$ be its transpose. Writing (41) in matrix form, we have

$$\boldsymbol{\psi}(\theta,t)^T = \boldsymbol{\phi}(\theta)\boldsymbol{r}\boldsymbol{\psi}(\theta,t-1)^T \tag{42}$$

with the initial condition

$$\boldsymbol{\psi}(\theta,1)^T = \boldsymbol{\phi}(\theta)\mathbf{1}^T,$$

where $\mathbf{1}^T$ is the column vector with all its elements being one. Let π_i be the probability of $x(0)$ being at state i and also let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$. Thus,

$$E(e^{\theta A(0,t)}) = \boldsymbol{\pi}\boldsymbol{\psi}(\theta,t)^T = \boldsymbol{\pi}(\boldsymbol{\phi}(\theta)\boldsymbol{r})^{t-1}\boldsymbol{\phi}(\theta)\mathbf{1}^T. \tag{43}$$

Since $\mathrm{sp}(\boldsymbol{\phi}(\theta)\boldsymbol{r})$ is the spectral radius of the matrix $\boldsymbol{\phi}(\theta)\boldsymbol{r}$, for every $\epsilon > 0$ there exists a constant $\sigma_\epsilon(\theta)$ such that every element of the matrix $(\boldsymbol{\phi}(\theta)\boldsymbol{r})^t$ is bounded above by $\sigma_\epsilon(\theta)(\mathrm{sp}(\boldsymbol{\phi}(\theta)\boldsymbol{r}) + \epsilon)^t$ (see [24], Corollary 5.6.13). In conjunction with (43), one can easily show that $a^*(\theta) \leq (1/\theta)\log\mathrm{sp}(\boldsymbol{\phi}(\theta)\boldsymbol{r})$.

If we also assume that the Markov process $x(t)$ with the transition matrix $\boldsymbol{r}$ is irreducible and aperiodic, then the matrix $\boldsymbol{r}$ is *primitive*, i.e., $\boldsymbol{r}^n > 0$ for some $n \geq 1$. Observing that the matrix $\boldsymbol{\phi}(\theta)$ is a diagonal matrix with positive diagonal elements, it is easy to see that the matrix $\boldsymbol{\phi}(\theta)\boldsymbol{r}$ is also primitive. From (43) and the Perron-Frobenius theorem ([24], Theorem 8.5.1), i.e.,

$$\lim_{t\to\infty} [\boldsymbol{\phi}(\theta)\boldsymbol{r}/\mathrm{sp}(\boldsymbol{\phi}(\theta)\boldsymbol{r})]^t = L(\theta) > 0$$

for some constant matrix $L(\theta)$, it follows that the MER is $(1/\theta)\log\mathrm{sp}(\boldsymbol{\phi}(\theta)\boldsymbol{r})$. □

If the modulating process is a two-state Markov chain, then the spectral radius of the matrix $\boldsymbol{\phi}(\theta)\boldsymbol{r}$ can be computed easily (see [24], pp. 39) and

$$a^*(\theta) = \frac{1}{\theta}\log\Big(\frac{r_{11}\phi_1(\theta) + r_{22}\phi_2(\theta) + \sqrt{(r_{11}\phi_1(\theta) - r_{22}\phi_2(\theta))^2 + 4r_{12}r_{21}\phi_1(\theta)\phi_2(\theta)}}{2}\Big). \tag{44}$$

For the usual voice model [37], one has in particular a constant number of arrivals, ν, at state 2 and no arrivals at state 1. Then $\phi_2(\theta) = \exp(\nu\theta)$ and $\phi_1(\theta) = 1$ and the MER

$$a^*(\theta) = \frac{1}{\theta}\log\Big(\frac{r_{11} + r_{22}\exp(\nu\theta) + \sqrt{(r_{11} + r_{22}\exp(\nu\theta))^2 - 4(r_{11} + r_{22} - 1)\exp(\nu\theta)}}{2}\Big). \tag{45}$$

Our definitions for MERs in (33) can also be easily extended to continuous-time models. For instance, we consider the two-state Markov modulated fluid process. The transition rate from state 1 to state 2 is λ and the transition rate from state 2 to state 1 is μ. Assume that there are no arrivals at state 1 and that the arrivals at state 2 is a constant rate process with rate ν. Using the backward equation as in Example 3.3, one can easily show that

$$a^*(\theta) = \frac{\theta\nu - \mu - \lambda + \sqrt{(\theta\nu - \mu + \lambda)^2 + 4\lambda\mu}}{2\theta}. \tag{46}$$

The MER in (46) is then the same as the effective bandwidth $\alpha(\zeta)$ in Gibbens and Hunt [20] with $\zeta = -\theta$.

Now we discuss some properties of MEPs and MERs. In Lemma 3.4 below, we establish bounds for the MEP and the MER of a superposition of K independent processes. The proof is direct.

Lemma 3.4 *Let $a(t) = \sum_{k=1}^{K} a_k(t)$ be a superposition of K independent processes. Then*

(i) $A^*(\theta, t) \le \sum_{k=1}^{K} A_k^*(\theta, t)$ *and* $a^*(\theta) \le \sum_{k=1}^{K} a_k^*(\theta)$.

(ii) *If, furthermore, $a_k(t)$, $k = 1, \ldots, K$ satisfy conditions $[C1-3]$, then $a(t)$ also satisfies conditions $[C1-3]$ with $a^*(\theta) = \sum_{k=1}^{K} a_k^*(\theta)$.*

If these K processes are not independent, a general bound can be obtained by Hölder's inequality (suggested by Joy Thomas). Note that for $m_k > 1$, $k = 1, \ldots, K$, $\sum_{k=1}^{K} (1/m_k) = 1$,

$$Ee^{\theta \sum_{k=1}^{K} A_k(t_1,t_2)} \le \prod_{k=1}^{K} (Ee^{\theta m_k A_k(t_1,t_2)})^{1/m_k}. \tag{47}$$

It then follows that

$$a^*(\theta) \le \inf_{\sum_{k=1}^{K} (1/m_k)=1} \sum_{k=1}^{K} a_k^*(m_k\theta). \tag{48}$$

As we shall prove in Lemma 3.5, $a_k^*(\theta)$'s are increasing in θ. The general bound in (48) is not as tight as that in Lemma 3.4. However, this general bound cannot be improved without any further assumptions. Consider the case that $a_k(t)$'s are identical, i.e., $a(t) = Ka_1(t)$. Then $a^*(\theta) = Ka_1^*(K\theta)$ which is the same as the right hand side of (48), taking into account the convexity of $\theta a_1^*(\theta)$.

We note that the process $a(t)$ is associated if $a_k(t)$, $k = 1, \ldots, K$ are associated. This follows from the fact that independent random variables are associated.

In the following lemma, we establish monotonicity results and bounds for MEPs and MERs. Define the essential supremum of a random variable X, denoted as $\|X\|_\infty$, to be the greatest lower bound of the set $\{x : \text{Prob}(X > x) = 0\}$ (see [36]), i.e.,

$$\|X\|_\infty = \inf\{x : \text{Prob}(X > x) = 0\}.$$

Thus, $\mathrm{Prob}(X \le \|X\|_\infty) = 1$ and $EX \le \|X\|_\infty$. Since the function $\exp(\theta x)$, $0 < \theta < \infty$, is strictly increasing and continuous in x, we have that $\|\exp(\theta X)\|_\infty = \exp(\theta \|X\|_\infty)$. Moreover, it is easy to verify that

$$\|X_1\|_\infty + \|X_2\|_\infty \ge \|X_1 + X_2\|_\infty.$$

Lemma 3.5 *MEPs are increasing in θ, i.e., for $0 < \theta_1 \le \theta_2 < \infty$, $A^*(\theta_1, t) \le A^*(\theta_2, t)$ for all t. Moreover, for $0 < \theta < \infty$,*

$$\sup_{s \ge 0} EA(s, s+t) \le A^*(\theta, t) \le \sup_{s \ge 0} \|A(s, s+t)\|_\infty. \tag{49}$$

As a direct consequence, MERs are increasing in θ and

$$\inf_{t \ge 1} \frac{1}{t} \sup_{s \ge 0} EA(s, s+t) \le a^*(\theta) \le \inf_{t \ge 1} \frac{1}{t} \sup_{s \ge 0} \|A(s, s+t)\|_\infty. \tag{50}$$

If $a(t)$ is stationary, then $\inf_{t \ge 1} \frac{1}{t} \sup_{s \ge 0} EA(s, s+t) = Ea(0)$. The lower bound in (50) implies that MERs are not less than the corresponding average rate.

Proof. From Jensen's inequality, it follows that for $0 < \theta_1 \le \theta_2 < \infty$

$$Ee^{\theta_2 A(s,s+t)} = Ee^{(\theta_2/\theta_1)\theta_1 A(s,s+t)} \ge (Ee^{\theta_1 A(s,s+t)})^{\theta_2/\theta_1}. \tag{51}$$

Taking the log function on both sides yields

$$\frac{1}{\theta_1} \log Ee^{\theta_1 A(s,s+t)} \le \frac{1}{\theta_2} \log Ee^{\theta_2 A(s,s+t)}.$$

Thus, we have $A^*(\theta_1, t) \le A^*(\theta_2, t)$. The first inequality in (49) also follows immediately from Jensen's inequality. For the second inequality in (49), observe that

$$Ee^{\theta A(s,s+t)} \le \|e^{\theta A(s,s+t)}\|_\infty = e^{\|\theta A(s,s+t)\|_\infty}.$$

Observe that both $\sup_{s \ge 0} EA(s, s+t)$ and $\sup_{s \ge 0} \|A(s, s+t)\|_\infty$ are subadditive in t. Thus, the inequalities in (50) hold. $\square$

In the following lemma, we show that both the upper and lower bounds for MEPs can be reached if $a(t)$ is bounded. Moreover, under the same condition, the MEP $A^*(\theta, t)$ and the MER $a^*(\theta)$ are continuous for $0 < \theta < \infty$. The proof is given in Appendix A.

Lemma 3.6 *If $a(t)$ is bounded, i.e., $a(t) \le M$ for some constant $M < \infty$ and for all t, then*

(i) *the upper bound of the MEP can be reached by letting* $\theta \to \infty$*, i.e.,*

$$\lim_{\theta\to\infty} A^*(\theta, t) = \sup_{s\geq 0} \|A(s, s+t)\|_\infty,$$

(ii) *the lower bound of the MEP can be reached by letting* $\theta \to 0$*, i.e.,*

$$\lim_{\theta\to 0} A^*(\theta, t) = \sup_{s\geq 0} EA(s, s+t),$$

(iii) $A^*(\theta, t)$ *and* $a^*(\theta)$ *are continuous for all* $0 < \theta < \infty$.

From Lemma 3.6, it follows that the MEP $A^*(\theta, t)$ of a bounded process $a(t)$ is continuous for $0 \leq \theta \leq \infty$ if one defines $A^*(0, t) = \sup_{s\geq 0} EA(s, s+t)$ and $A^*(\infty, t) = \sup_{s\geq 0} \|A(s, s+t)\|_\infty$. We note that the conditions for $a^*(\theta)$ to be continuous at $\theta = 0$ and $\theta = \infty$ are in general more restrictive than the boundedness of $a(t)$. These conditions won't be pursued here.

As in the previous section, we shall use a lower case letter to denote a stochastic process, e.g., $a(t)$ and the corresponding upper case letter to denote its partial sums , e.g., $A(t_1, t_2) = \sum_{t=t_1}^{t_2-1} a(t)$. A superscript $*$ on the corresponding upper (lower) case letter will denote the MEP (MER) of that process, e.g., $A^*(\theta, t)$ $(a^*(\theta))$.

3.2 A single queue with multiple classes of customers

In this section, we consider a discrete-time queue with K classes of customers. The service requirements of these K class customers are assumed to be one unit of time. Let $a_k(t)$, $k = 1, \ldots, K$, be the number of class k arrivals at time t and $a(t) = \sum_{k=1}^K a_k(t)$ be the total number of arrivals at time t. We assume that these K arrival processes are *independent*. Denote $q(t)$ as the number of customers in the queue at time t. Assume that the buffer size is infinite and that the server can serve c customers per unit of time. The constant c will be referred to as the capacity of the server. Analogous to §2.2, under a work-conserving policy the queue is governed by Lindley's equation in (5). Furthermore, we assume that the queue is empty at time 0.

Let $A_k(t_1, t_2) = \sum_{t=t_1}^{t_2-1} a_k(t)$ be the number of class k arrivals in $[t_1, t_2)$ and $A_k^*(\theta, t)$ be its MEP with MER $a_k^*(\theta)$. We use the notations without the subscript k to denote the corresponding definitions for the superposition of these K independent processes. For the departure processes, we use the letter b or B to denote the corresponding quantities.

In the following lemma, we establish an input-output relation for a single queue. We show that if the arrival process of each class has a linear EP, $\hat{a}_k(\theta)t + \hat{\sigma}_k(\theta)$, and the total envelope rate is less than the capacity, i.e., $\sum_{k=1}^{K} \hat{a}_k(\theta) < c$, then (i) the queue length is bounded exponentially with respect to θ, (ii) there exists a linear EP of the departure process which can be represented as a function of the linear EP of the input process and (iii) the virtual delay at time t (the workload at time t) is bounded exponentially with respect to θc if the scheduling policy is First Come First Served (FCFS).

Lemma 3.7 *Suppose that $\hat{a}_k(\theta)t + \hat{\sigma}_k(\theta)$ is an EP of $a_k(t)$, i.e., $A_k^*(\theta, t) \le \hat{a}_k(\theta)t + \hat{\sigma}_k(\theta)$. Let $\hat{a}(\theta) = \sum_{k=1}^{K} \hat{a}_k(\theta)$ and $\hat{\sigma}(\theta) = \sum_{k=1}^{K} \hat{\sigma}_k(\theta)$. Also let $B_S^*(\theta, t)$ be the MEP of $\sum_{k \in S} b_k(t)$, where S is a subset of $\{1, \dots, K\}$. If $\hat{a}(\theta) < c$, then $q(t)$ is bounded exponentially with respect to θ, and there exists a constant $\beta(\theta) < \infty$ such that for all t,*

$$Prob(q(t) \ge x) \le \beta(\theta) e^{-\theta x} \tag{52}$$

$$B_S^*(\theta, t) \le (\sum_{k \in S} \hat{a}_k(\theta))t + \frac{1}{\theta} \log \beta(\theta), \tag{53}$$

where

$$\beta(\theta) = e^{\theta \hat{\sigma}(\theta)} (1 - e^{\theta(\hat{a}(\theta) - c)})^{-1}.$$

If the scheduling policy is FCFS, then the virtual delay at time t, denoted as $v(t)$, is bounded exponentially with respect to θc and

$$Prob(v(t) \ge x) \le e^{\theta \hat{a}(\theta)} \beta(\theta) e^{-\theta c(x-1)}. \tag{54}$$

Proof. Expanding (5) recursively yields

$$q(t) = \max\Big[0, a(t-1) - c, a(t-1) + a(t-2) - 2c, \dots, a(t-1) + a(t-2) + \dots a(0) - tc\Big]. \tag{55}$$

Using the inequality that $\max(x_1, x_2) \le x_1 + x_2$ for $x_1, x_2 \ge 0$, we have

$$E e^{\theta q(t)} \le \sum_{s=0}^{t} E e^{\theta(A(t-s,t) - sc)}. \tag{56}$$

From (31) and Lemma 3.4, it follows that $E \exp(\theta A(t-s, t)) \le \exp(\theta \hat{a}(\theta)s + \theta \hat{\sigma}(\theta))$. In conjunction with (56),

$$E e^{\theta q(t)} \le e^{\theta \hat{\sigma}(\theta)} \sum_{s=0}^{t} e^{\theta s(\hat{a}(\theta) - c)} \le e^{\theta \hat{\sigma}(\theta)} \sum_{s=0}^{\infty} e^{\theta s(\hat{a}(\theta) - c)} = \beta(\theta) \tag{57}$$

if $\hat{a}(\theta) < c$. Applying Chernoff's bound yields

$$\mathrm{Prob}(q(t) \geq x) \leq e^{-\theta x} E e^{\theta q(t)} \leq \beta(\theta) e^{-\theta x}.$$

This completes the argument for the queue length.

For the departure processes, observe that the number of class k departures in $[t_1, t_2)$ is not greater than the sum of the number of class k arrivals in $[t_1, t_2)$ and the number of class k customers in the queue at time t_1. Thus, we have

$$\sum_{k \in S} B_k(t_1, t_2) \leq \sum_{k \in S} A_k(t_1, t_2) + q(t_1) \tag{58}$$

for any subset S of $\{1, \ldots, K\}$. From (55), it follows that

$$\sum_{k \in S} B_k(t_1, t_2) \leq \max_{0 \leq s \leq t_1} \Big[\sum_{k \in S} A_k(t_1 - s, t_2) + \sum_{k \notin S} A_k(t_1 - s, t_1) - sc \Big]. \tag{59}$$

Using an argument similar to that for the queue length and the independence assumption of arrival processes, one can easily show that

$$E \exp(\theta \sum_{k \in S} B_k(t_1, t_2)) \leq \exp(\theta \sum_{k \in S} \hat{a}_k(\theta)(t_2 - t_1)) \beta(\theta).$$

Taking the log function on both sides completes the argument for the departure processes.

If the scheduling policy is FCFS, then the virtual delay of a customer that arrives at time t is bounded above by $\lceil (q(t) + a(t))/c \rceil$. Note that

$$\mathrm{Prob}(\lceil (q(t) + a(t))/c \rceil \geq x) \leq \mathrm{Prob}(q(t) + a(t) \geq c(x-1)). \tag{60}$$

Now $q(t) + a(t)$ in (60) is a special case of (58) when taking $t_1 = t$, $t_2 = t + 1$ and $S = \{1, 2, \ldots, K\}$. $\square$

We note that the K departure processes are in general not independent though the K arrival processes are independent. Moreover, the bounding processes for the departure processes obtained by (58) are in general not independent since the random variable $q(t)$ appears in the right hand side of (58) for each class. However, if the queue length is always bounded above by a constant q, one could obtain independent bounding processes for the departure processes by replacing $q(t)$ with q in (58). If, furthermore, the delay of each customer is bounded above by a constant d, one can use the property derived in a deterministic queue (cf. Lemma 2.6) to establish that $B_k(t_1, t_2) \leq A_k(t_1 - d, t_2)$. Now the bounding processes $A_k(t_1 - d, t_2), k = 1, 2, \ldots, K$, are also independent if the arrival processes $a_k(t), k = 1, 2, \ldots, K$, are independent. These independent bounding processes have been used in Kurose [30] for stochastic networks with deterministic bounded delays.

Also, we note that inequalities similar to (52) for queues with renewal inputs, i.e., $GI/GI/1$ queues, were reported in the literature (see [27, 35, 38]).

From the relation between the MER and the class of linear EPs in (34-35), the theorem below, stating the input-output relation of MERs and the boundedness of queue length and virtual delay, follows as a direct consequence of Lemma 3.7.

Theorem 3.8 *If the sum of MER of the K independent processes is less than the capacity, i.e., $\sum_{k=1}^{K} a_k^*(\theta) < c$, then the MER of the departure process is bounded above by the MER of the corresponding arrival process, i.e., $b_k^*(\theta) \leq a_k^*(\theta)$ for all k. Moreover, the queue length can be bounded exponentially with respect to θ. If the scheduling policy is FCFS, then the virtual delay can be bounded exponentially with respect to θc.*

Since $a^*(\theta)$ is increasing in θ (Lemma 3.5), it is of interest to study the largest θ that satisfies $a^*(\theta) < c$. Let

$$\theta^* = \sup\{\theta : a^*(\theta) < c\}. \tag{61}$$

Suppose $a_k(t)$ is a two-state Markov modulated process with $a_k^*(\theta)$ described in (45). Then θ^* is the solution of the following equation

$$\exp(\theta c) = \prod_{k=1}^{K} \Big(\frac{r_{11}^{(k)} + r_{22}^{(k)} \exp(\nu_k \theta) + \sqrt{(r_{11}^{(k)} + r_{22}^{(k)} \exp(\nu_k \theta))^2 - 4(r_{11}^{(k)} + r_{22}^{(k)} - 1)\exp(\nu_k \theta)}}{2} \Big). \tag{62}$$

We note that (62) is the same as (13) in Sohraby [37], which was obtained by a spectral decomposition method. Based on an asymptotic expansion, Sohraby further obtained an approximation for θ^* and showed that it is consistent with the result in [1] when all the K arrival processes are identically distributed.

We now show a converse statement to Theorem 3.8.

Theorem 3.9 *(i) If the MER of the input process is larger than the capacity, i.e., $a^*(\theta) > c$, then the queue length cannot be bounded exponentially with respect to θ, i.e., there does not exist a constant $d < \infty$ such that $(E \exp(\theta q(t)))^{1/\theta} \leq d$ for all t.*

(ii)If, furthermore, $a_k(t)$, $k = 1, \ldots, K$, satisfy conditions $[C1-3]$ and θ^* in (61) is positive and finite, i.e., $0 < \theta^* < \infty$, then the queue length process $\{q(t), t \geq 0\}$ converges in distribution

to a finite random variable $q(\infty)$ that satisfies

$$\lim_{x\to\infty} \frac{-\log Prob(q(\infty) \geq x)}{x} = \theta^*. \tag{63}$$

Proof. (i) We will prove the first part of the theorem by contradiction. Assume that $E\exp(\theta q(t)) \leq d^\theta < \infty$ for all t. From (55), it follows that $q(s+t) \geq A(s, s+t) - tc$. Thus, $E\exp(\theta(A(s, s+t) - tc)) < \infty$ for all s and t. This in turns implies that $a^*(\theta) \leq c$ and we reach a contradiction.

(ii) From Lemma 3.4(ii), it follows that $a^*(\theta) = \sum_{k=1}^K a_k^*(\theta)$ and $a^*(\theta)$ also satisfies conditions $[C1-3]$. Thus, $a^*(\theta)$ is strictly increasing, and θ^* is the unique solution of $a^*(\theta) = c$. Since $0 < \theta^* < \infty$, we have from Lemma 3.5 that $Ea(0) < a^*(\theta^*) = c$. Observe from the stationarity of $a(t)$ and (55) that the queue length process $\{q(t), t \geq 0\}$ is a stochastically increasing sequence if $q(0) = 0$ (cf. Loynes's construction in [31, 3]). Recall that the stochastic ordering $X \leq_{st} Y$ if $\mathrm{Prob}(X \geq x) \leq \mathrm{Prob}(Y \geq x)$ for all x. Thus, $\{q(t), t \geq 0\}$ converges in distribution to a finite random variable $q(\infty)$. To show (63), observe from Theorem 3.8 that

$$\limsup_{x\to\infty} \frac{\log \mathrm{Prob}(q(t) \geq x)}{x} \leq -\theta \tag{64}$$

for all t and x if $a^*(\theta) < c$. Thus, it suffices to show the lower bound. Since $q(t) \leq_{st} q(\infty)$, $\mathrm{Prob}(q(\infty) \geq x) \geq \mathrm{Prob}(q(t) \geq x) \geq \mathrm{Prob}(A(0,t) \geq ct + x)$ for all t. Letting $v = c + (x/t)$ and applying the lower bound of the Gärtner-Ellis theorem [8], one has

$$\liminf_{x\to\infty} \frac{\log \mathrm{Prob}(q(\infty) \geq x)}{x} \geq \liminf_{t\to\infty} \frac{\log \mathrm{Prob}(A(0,t) \geq vt)}{(v-c)t} \geq -\frac{I(v)}{v-c}, \tag{65}$$

where $I(v)$ is defined in (36). Optimizing v over all possible values yields

$$\liminf_{x\to\infty} \frac{\log \mathrm{Prob}(q(\infty) \geq x)}{x} \geq -\inf_{v \geq c} \frac{I(v)}{v-c}. \tag{66}$$

The proof is then completed if the right hand side of (66) is shown to be $-\theta^*$. We will follow the argument used in [29] Lemma 1. Let $J(\theta) = \theta a^*(\theta)$. Then $I(v)$ and $J(\theta)$ are convex conjugates [2]. Since we assume by $[C3]$ that $J(\theta)$ is strictly convex and differentiable, $I(v)$ and $J(\theta)$ actually forms a pair of Legendre transformation, i.e., $I(v)$ is also strictly convex and differentiable and $J(\theta)$ has the representation

$$J(\theta) = \theta {I'}^{-1}(\theta) - I({I'}^{-1}(\theta)), \tag{67}$$

where $I'^{-1}(\theta)$ is the inverse function of $I'(v)$. Since θ^* is the unique solution of $a^*(\theta) = c$, θ^* is the unique solution of

$$I'^{-1}(\theta) - \frac{I(I'^{-1}(\theta))}{\theta} = c. \tag{68}$$

Define the function $g(v) = v - I(v)/I'(v)$. From the strict convexity of $I(v)$, it follows that $g(v)$ is strictly increasing. Thus we can define g^{-1} as the inverse function of g. Note that $g^{-1}(c)$ is the solution of the equation $c = v - I(v)/I'(v)$ and that

$$\inf_{v \geq c} \frac{I(v)}{v - c} = I'(g^{-1}(c)). \tag{69}$$

It is easy to verify that $I'(g^{-1}(c))$ is indeed a solution of (68) and thus $\inf_{v \geq c} I(v)/(v-c) = \theta^*$.
$\square$

In particular, we consider the arrival process as a superposition of K independent continuous-time two-state Markov modulated fluid processes with the MERs $a_k^*(\theta)$ in (46). It is easy to see that

$$a^*(\theta) = \sum_{k=1}^{K} \frac{\theta\nu_k - \mu_k - \lambda_k + \sqrt{(\theta\nu_k - \mu_k + \lambda_k)^2 + 4\lambda_k\mu_k}}{2\theta} \tag{70}$$

for all $0 < \theta < \infty$ and thus conditions $[C1-3]$ are satisfied. As an application of Theorems 3.9(ii), we recover the result for the effective bandwidth in Gibbens and Hunt, Theorem 1 [20]. Note that θ^* is the solution of the equation

$$c = \sum_{k=1}^{K} \frac{\theta\nu_k - \mu_k - \lambda_k + \sqrt{(\theta\nu_k - \mu_k + \lambda_k)^2 + 4\lambda_k\mu_k}}{2\theta}. \tag{71}$$

The equation (71) has been reported by Guérin, Ahmadi and Naghshineh ([23], (7)) for an approximation of the tail distribution of the queue length.

To extend our result for delay, define the distribution of the stationary delay z as follows:

$$\text{Prob}(z \geq x) = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \mathbf{1}_{\{z_m \geq x\}}, \tag{72}$$

where z_m is the delay of the m^{th} customer that arrives at the queue.

Corollary 3.10 *If (i) $a(t)$ satisfy conditions $[C1-3]$, (ii) $1 \leq a(t) \leq M < \infty$, and (iii) θ^* in (61) is positive and finite, then under the FCFS policy*

$$\lim_{x\to\infty} \frac{-\log Prob(z \geq x)}{x} = \theta^* c. \tag{73}$$

Proof. Note that under the FCFS policy

$$\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n}\mathbf{1}_{\{z_m\geq x\}}=\frac{\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\sum_{l=1}^{a(s)}\mathbf{1}_{\{\lceil (q(s)+l)/c\rceil\geq x\}}}{\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}a(s)}. \tag{74}$$

Since $q(s)\leq q(s)+l\leq q(s)+a(s)$, it then follows from stationarity and ergodicity (cf. Campbell-Little-Mecke formula in [3]) that

$$\frac{Ea(0)\mathbf{1}_{\{\lceil q_s(0)/c\rceil\geq x\}}}{Ea(0)}\leq \mathrm{Prob}(z\geq x)\leq \frac{Ea(0)\mathbf{1}_{\{\lceil (q_s(0)+a(0))/c\rceil\geq x\}}}{Ea(0)}, \tag{75}$$

where $q_s(0)$ is the stationary version of $q(t)$ at time 0, i.e., $q_s(0)=_{st}q(\infty)$. The rest of the proof then follows from Theorem 3.9(ii) using the assumption $1\leq a(0)\leq M$. □

In the next section, we will use the input-output relation in Lemma 3.7 to study the stability of acyclic networks in which we need the notion of splitting departure processes.

3.3 Marked point processes and acyclic networks

Our objective in this section is to extend the single queue result to acyclic networks. Our approach is based on the notion of marked point processes. A discrete-time marked point process $\{(\tau(n),p(n)),n=0,1,2,\ldots\}$ is a sequence of random vectors. The process $\{\tau(n)\}$ is called the arrival process with $\tau(n)$ being the arrival epoch of the n^{th} customer. We will assume that $\tau(n)$ is increasing and that $\tau(n)\to\infty$ *a.s.* as $n\to\infty$. The random variable $p(n)$ is called the mark associated with the arrival epoch $\tau(n)$. For instance, the mark $p(n)$ could be the service requirement of the n^{th} arrival. If $p(n)$ is an indicator function, it could be viewed as a routing variable (see Section 2.3). Let

$$b(t)=\sum_{n=0}^{\infty}\mathbf{1}_{\{\tau(n)=t\}}.$$

Then $b(t)$ is the number of arrivals at time t. Let

$$a(t)=\sum_{n=0}^{\infty}\mathbf{1}_{\{\tau(n)=t\}}p(n).$$

If $p(n)$'s are service requirements, then $a(t)$ is the total amount of work that arrives at time t. On the other hand, if $p(n)$'s are indicator random variables (e.g., routing variables), then $a(t)$ is a thinning process of $b(t)$ and can be viewed as the number of customers that are routed to a particular queue at time t.

In the following, we establish the "input-output" relation between the arrival process and the marked process.

Lemma 3.11 *(i) If the two sequences $\{\tau(n), n \geq 0\}$ and $\{p(n), n \geq 0\}$ are independent, and there exist two linear EPs of $b(t)$ and $p(n)$ as follows:*

$$B^*(\theta, t) \leq \hat{b}(\theta)t + \hat{\delta}(\theta) \;\; \forall\, 0 < \theta < \infty,$$
$$P^*(\theta, m) \leq \hat{p}(\theta)m + \hat{\eta}(\theta) \;\; \forall\, 0 < \theta < \infty,$$

then there exists a linear EP of the marked process $a(t)$ as follows:

$$A^*(\theta, t) \leq \hat{p}(\theta)\hat{b}(\theta\hat{p}(\theta))t + \hat{\eta}(\theta) + \hat{p}(\theta)\hat{\delta}(\theta\hat{p}(\theta)) \;\; \forall\, 0 < \theta < \infty.$$

As a direct consequence, the MER of the marked point process, $a^*(\theta)$, is bounded above by $p^*(\theta)b^*(\theta p^*(\theta))$, where $b^*(\theta)$ and $p^*(\theta)$ are the MERs of the arrival process and the marking process respectively.

(ii) If, furthermore, both $b(t)$ and $p(n)$ satisfies conditions $[C1-3]$, then $a(t)$ also satisfies conditions $[C1-3]$ with $a^*(\theta) = p^*(\theta)b^*(\theta p^*(\theta))$.

We note that we have implicitly assumed that $p^*(\theta) < \infty$ in Lemma 3.11. If $p^*(\theta) = \infty$, then we have the trivial inequality $a^*(\theta) \leq \infty$.

Proof. (i) Let $n_f = \inf\{n : \tau(n) \geq t_1\}$ and $n_l = \inf\{n \geq n_f : \tau(n) \geq t_2\}$. Since $\tau(n) \to \infty$ *a.s.* as $n \to \infty$, both n_f and n_l are finite random variables. In other words, n_f is the identity of the first customer that arrives after $t_1 - 1$ and n_l is the identity of the first customer that arrives after $t_2 - 1$. Thus, $B(t_1, t_2) = n_l - n_f$. Since $\{\tau(n)\}$ and $\{p(n)\}$ are independent,

$$\begin{aligned}
Ee^{\theta A(t_1,t_2)} &= \sum_{m_1=0}^{\infty}\sum_{m_2=m_1}^{\infty} Ee^{\theta\sum_{n=m_1}^{m_2-1}p(n)}\mathrm{Prob}(n_f = m_1, n_l = m_2) \\
&\leq \sum_{m_1=0}^{\infty}\sum_{m_2=m_1}^{\infty} e^{\theta(\hat{p}(\theta)(m_2-m_1)+\hat{\eta}(\theta))}\mathrm{Prob}(n_f = m_1, n_l = m_2) \\
&= \sum_{m_1=0}^{\infty}\sum_{m=0}^{\infty} e^{\theta(\hat{p}(\theta)m+\hat{\eta}(\theta))}\mathrm{Prob}(n_f = m_1, B(t_1,t_2) = m)
\end{aligned}$$

Interchanging the sums yields

$$Ee^{\theta A(t_1,t_2)} \leq e^{\theta\hat{\eta}(\theta)}Ee^{\theta\hat{p}(\theta)B(t_1,t_2)}. \tag{76}$$

Replacing the expectation in (76) by the EP of $B(t_1, t_2)$ and taking the log function on both sides completes the derivation.

(ii) It follows directly from (37) and the argument for (i). For stationarity and ergodicity of marked point processes, we refer to [3]. $\square$

We note that the independence of the arrival process and the marking process is crucial. As we mentioned earlier, the departure processes of different classes of customers from a common queue are not independent. Thus, in order to use the input-output relation in a network, the departure processes of different classes from a common queue cannot be the arrival processes of another queue. Networks with this property are known as acyclic networks (see figure 4). To be precise, consider the multiclass feedforward network in Section 2.3. Construct a directed graph with all the queues in the feedforward network being its nodes (excluding the router). Add an arc between queue i and queue j $(i < j)$ whenever there is a class of customers that are routed from queue i to queue j. Then the network is acyclic if there is at most one path from queue i to queue j $(i < j)$ in the directed graph. This implies that the input in front of each queue can be represented as a superposition of independent arrival processes. Thus, one could apply Lemmas 3.4, 3.7 and 3.11 inductively to obtain bounds for the tail distribution of queue length at each queue as well as the linear EPs of the input process and the departure process at each queue.

Analogous to the notations in Section 2.3, let $a^*_{i,k}(\theta)$ and $p^*_{i,j,k}(\theta)$ be the MER of the arrival process of class k customer to queue i and the MER of the routing sequence of class k customer from queue i to queue j ($i = 0$ for external arrivals). Let $S_{j,k}$ denote the set of queues from which there are class k customers routed to queue j. In an acyclic network, the set $S_{j,k}$ contains at most one element. Let $\rho_{0,k}(\theta) = a^*_{0,k}(\theta), k = 1, \ldots, K$. Since an acyclic network is feedforward, we can define recursively for each k

$$\rho_{j,k}(\theta) = \sum_{i \in S_{j,k}} \rho_{i,k}(\theta p^*_{i,j,k}(\theta)) p^*_{i,j,k}(\theta), \quad j = 1, \ldots, I. \tag{77}$$

Theorem 3.12 *In an acyclic network, if $\sum_{k=1}^{K} \rho_{j,k}(\theta) < c_j$ for all $j = 1, \ldots, I$, then the queue length of each queue can be bounded exponentially with respect to θ. If the scheduling policy is FCFS at each queue, then the virtual delay of a customer that arrives at queue i at time t can also be bounded exponentially with respect to θc_i.*

Proof. Using an argument similar to that used in Section 2.3 and Lemma 3.11, one can easily show inductively that $\rho_{j,k}(\theta) \geq a^*_{j,k}(\theta)$. The rest of the proof of Theorem 3.12 then follows from Theorem 3.8. $\square$

We note that there are other networks that can be analyzed by our method. For instance, the intree network in figure 5 is not an acyclic network since there are two paths from queue 1 to queue 2. However, using (53) we are still able to obtain a linear EP of the superposition of the departure processes from queue 1. Thus, by viewing the superposition of the departure processes from queue 1 as a single composite process, the input process at queue 2 can be represented as a superposition of independent processes with known linear EPs. Using Lemma 3.7, one could obtain a bound for the tail distribution of the queue length at queue 2, as well as other desired information.

To extend from the fixed service requirements to general service requirements, one can consider the virtual waiting processes instead of the queue length processes. Since the virtual waiting process can be approximated by the queue length process subject to batch arrivals, similar results to (52) and (54) in Lemma 3.7 can be derived by replacing the queue length $q(t)$ by the work load $v(t)$ and treating $a(t)$ as a marked point process with the marks representing the service requirements. However, unlike the deterministic queues with bounded delays in the previous section for deterministic networks, we do not have the input-output relation as in (53) and the result might not able to be extended to acyclic networks.

3.4 Single class networks with nonfeedforward routing

In this section, we consider the nonfeedforward network in Section 2.4. As in the acyclic networks, we assume that the sequences of routing random variables $\{p_{i,j}(n)\}$ and the external arrival process $\{a_0(t)\}$ are independent. We further assume that the sequence of the routing variables $\{(p_{i,1}(n), \ldots, p_{i,I}(n))\}$ are i.i.d. random vectors with the means $(\bar{p}_{i,1}, \ldots, \bar{p}_{i,I})$. Thus, the sequence $\{p_{i_1,j}(n), n = 0, 1, 2, \ldots\}$ and the sequence $\{p_{i_2,j}(n), n = 0, 1, 2, \ldots\}$ $(i_1 \neq i_2)$ form two *independent* sequences of i.i.d. Bernoulli random variables with means $\bar{p}_{i_1,j}$ and $\bar{p}_{i_2,j}$ respectively. This implies that

$$P^*_{i,j}(\theta, m) = m p^*_{i,j}(\theta)$$

with

$$p^*_{i,j}(\theta) = \frac{1}{\theta} \log(\bar{p}_{i,j} e^\theta + (1 - \bar{p}_{i,j})).$$

We note that $p_{i,j_1}(n)$ and $p_{i,j_2}(n)$ $(j_1 \neq j_2)$ are in general not independent.

Let $A^*_0(\theta, t)$ and $a^*_0(\theta)$ be the MEP and the MER of $a_0(t)$. In the following theorem, we show that every queue in the single class nonfeedforward network is bounded exponentially with respect to θ if a strong traffic condition similar to E_1 in Section 2.4 is satisfied.

Theorem 3.13 *If $\hat{a}_0(\theta)t + \hat{\sigma}_0(\theta)$ is a linear EP of the external arrival process $a_0(t)$ satisfying*

$$\hat{a}_0(\theta p^*_{0,j}(\theta))p^*_{0,j}(\theta) + \sum_{i=1}^{I} c_i p^*_{i,j}(\theta) < c_j \quad \text{for } j = 1, \ldots, I, \tag{78}$$

then for $j = 1, \ldots, I$, the queue length of the j^{th} queue at time t, $q_j(t)$, is bounded exponentially with respect to θ and

$$P(q_j(t) \geq x) \leq \beta_j(\theta)e^{-\theta x},$$

where

$$\beta_j(\theta) = \exp\left[\theta\hat{\sigma}_0(\theta p^*_{0,j}(\theta))p^*_{0,j}(\theta)\right]\Big(1 - \exp\left[\theta(\hat{a}_0(\theta p^*_{0,j}(\theta))p^*_{0,j}(\theta) + \sum_{i=1}^{I} c_i p^*_{i,j}(\theta) - c_j)\right]\Big)^{-1}.$$

As a direct consequence, the queue length is bounded exponentially with respect to θ if

$$a^*_0(\theta p^*_{0,j}(\theta))p^*_{0,j}(\theta) + \sum_{i=1}^{I} c_i p^*_{i,j}(\theta) < c_j \quad \text{for all } j. \tag{79}$$

Similar results hold for the virtual delay at each queue when the scheduling policy is FCFS.

Proof. Let $A_{i,j}(t_1, t_2)$, $i = 0, \ldots, I$, $j = 1, \ldots, I$, be the number of customers that are routed from queue i to queue j in the interval $[t_1, t_2)$ ($i = 0$ for external arrivals) with the corresponding MEP $A^*_{i,j}(\theta, t)$ and MER $a^*_{i,j}(\theta)$. Also let $A_j(t_1, t_2) = \sum_{i=0}^{I} A_{i,j}(t_1, t_2)$ be the total number of customers that arrives at queue j within the interval $[t_1, t_2)$ with the corresponding MEP $A^*_j(\theta, t)$ and MER $a^*_j(\theta)$. In order to use Lemma 3.7 to derive the desired result, we need to derive a linear EP for $A_j(t_1, t_2)$.

From Lemma 3.11, it follows that for $j = 1, \ldots, I$,

$$A^*_{0,j}(\theta, t) \leq \hat{a}_0(\theta p^*_{0,j}(\theta))p^*_{0,j}(\theta)t + \hat{\sigma}_0(\theta p^*_{0,j}(\theta))p^*_{0,j}(\theta). \tag{80}$$

Note that the number of customers that depart from queue i in the interval $[t_1, t_2)$ is bounded by $c_i(t_2 - t_1)$. Since we assume that the routing random variables are i.i.d., it then follows from a standard sample path argument (see [38, 35]) that

$$A_j(t_1, t_2) \leq_{st} A_{0,j}(t_1, t_2) + \sum_{n=1}^{c_1(t_2-t_1)} p_{1,j}(n) + \ldots + \sum_{n=1}^{c_I(t_2-t_1)} p_{I,j}(n), \tag{81}$$

where $\leq_{st}$ denotes the stochastic ordering as in the proof of Theorem 3.9(ii). Now the right-hand side of (81) is a sum of *independent* random variables. Using an argument similar to Lemma 3.11 and (80), one can show that

$$A_j^*(\theta,t) \leq \Big(\hat{a}_0(\theta p_{0,j}^*(\theta))p_{0,j}^*(\theta) + \sum_{i=1}^{I} c_i p_{i,j}^*(\theta)\Big)t + \hat{\sigma}_0(\theta p_{0,j}^*(\theta))p_{0,j}^*(\theta).$$

View the superposition process of all arrivals to queue j as a single class of customers. Applying Lemma 3.7 for a single class of customers completes the proof. □

Analogous to the notations in Section 2.4, let $\nu_j(\theta) = a_0^*(\theta p_{0,j}^*(\theta))p_{0,j}^*(\theta)$, $j = 1, \ldots, I$. From Lemma 3.11, $\nu_j(\theta)$ is an upper bound of the MER of the external arrivals to queue j. Using the matrix representation, let $\boldsymbol{\nu}(\theta) = [\nu_1(\theta), \ldots, \nu_I(\theta)]$, $\boldsymbol{c} = [c_1, \ldots, c_I]$ and $\boldsymbol{p}^*(\theta)$ be a matrix with $p_{i,j}^*(\theta)$ being the element in the i^{th} row and j^{th} column. The condition in (79) then corresponds to the polyhedral set defined in Section 2.4:

$$E_1(\theta) = \{\boldsymbol{c} : \boldsymbol{\nu}(\theta) < \boldsymbol{c}(\boldsymbol{\Gamma} - \boldsymbol{p}^*(\theta))\}, \tag{82}$$

where $\boldsymbol{\Gamma}$ is the identity matrix. If the spectral radius of the matrix $\boldsymbol{p}^*(\theta)$, denoted by $\mathrm{sp}(\boldsymbol{p}^*(\theta))$, is less than 1, then the matrix $\boldsymbol{\Gamma} - \boldsymbol{p}^*(\theta)$ is invertible and the traffic equation

$$\boldsymbol{\rho}(\theta) = \boldsymbol{\nu}(\theta) + \boldsymbol{\rho}(\theta)\boldsymbol{p}^*(\theta) \tag{83}$$

has the solution

$$\boldsymbol{\rho}(\theta) = (\rho_1(\theta), \ldots, \rho_I(\theta)) = \boldsymbol{\nu}(\theta)(\boldsymbol{\Gamma} - \boldsymbol{p}^*(\theta))^{-1}.$$

Consider the second polyhedral set:

$$E_2(\theta) = \{\boldsymbol{c} : \boldsymbol{\nu}(\theta)(\boldsymbol{\Gamma} - \boldsymbol{p}^*(\theta))^{-1} < \boldsymbol{c}\} \tag{84}$$

Analogous to the argument in Section 2.4, we have the following two properties between these two open polyhedral sets (if $\mathrm{sp}(\boldsymbol{p}^*(\theta)) < 1$).

(P1) $E_1(\theta) \subset E_2(\theta)$.

(P2) For every vector $\boldsymbol{c} \in E_2(\theta)$, there is a vector $\boldsymbol{c}^1 \in E_1(\theta)$ such that $\boldsymbol{c}^1 \leq \boldsymbol{c}$.

Theorem 3.14 *If $\rho_j(\theta) < c_j$ for all j and $sp(\boldsymbol{p}^*(\theta)) < 1$, then the total number of customers in the system can be bounded exponentially with respect to θ/I.*

Proof. Since for every $\boldsymbol{c}^2 \in E_2$, there exists $\boldsymbol{c}^1 \in E_1$ such that $\boldsymbol{c}^1 \leq \boldsymbol{c}^2$. From Theorem 3.13, the queue length at each queue is bounded exponentially with respect to θ when the system has capacity $\boldsymbol{c}^1$. Note that if there are non-integer component in $\boldsymbol{c}^1$, they should be interpreted as periodic sequences as in Remark 2.11. Using the Hölder's inequality in (47) yields the result that the total number of customers in the system is bounded exponentially with respect to θ/I. Applying the monotonicity result in Lemma 2.10, the total number of customers in the system with capacity $\boldsymbol{c}^2$ is then bounded exponentially with respect to θ/I. $\square$

As an application of Theorem 3.14, consider the case that the external arrivals $a_0(t)$ are i.i.d. random variables with mean $\bar{a}_0$. Suppose the moment generating function of $a_0(t)$, denoted as $\phi_0(\theta)$, is finite for *some* $\theta > 0$. Then $a^*(\theta) = 1/\theta \log \phi_0(\theta)$ and $\lim_{\theta \to 0} a_0^*(\theta) = \bar{a}_0$. Similarly, $\lim_{\theta \to 0} p_{i,j}^*(\theta) = \bar{p}_{i,j}$. Thus, if one is only interested in whether the sequence of distributions of the total number of customers in the network at time t is tight or not, the traffic equations $\rho_j(\theta) < c_j$ can be replaced by the traffic equations using average rates. This has been reported in [18] under similar moment conditions. Along this line, we have proposed a unified approach for the stability of generalized Jackson's networks in [10].

To stabilize a K-class nonfeedforward network, one can reserve a certain portion of the capacity at each queue to each class as discussed in Section 2.4. However, the problem of how one partitions the capacity is interesting and requires more numerical study.

4 Conclusions and future research

In this paper, we have proposed two new notions of traffic characterization: MER and MER with respect to θ. We have also developed a set of rules for network operations based on these two characterizations. These rules provide a method to answer two types of stability problem of queueing networks: (i) conditions for queueing networks that render bounded queue lengths and bounded delay for customers, and (ii) conditions for queueing networks in which the queue length distribution of a queue has an exponential tail with rate θ. For single class networks with nonfeedforward routing, we have provided a new method to establish stability results under the FCFS policy.

Recently, we have extended our theory in two directions: (i) large deviation and fast simulation, and (ii) stability of other networks. The connection with large deviation theory through Gärtner-Ellis theorem was first established in [29] using the Legendre transform. Along this line, we have extended the notion of envelope process with respect to θ in [11], where a fast simulation method for ATM intree networks is derived. The new method for the stability of nonfeedforward networks has been applied to generalized Jackson's networks in [10]. Another possible application is the stability of token rings with limited service. We note that Yaron and Sidi, in a recent paper [41], also considered exponential bounds as in Section 3. The key difference between our work and theirs is that we allow the bounds to be parameterized by θ, which in general renders tighter bounds and sometimes lower bounds.

Finally, we note that the notion of MER with respect to θ might be of practical importance in communication networks. In Section 3, we have shown that the MER with respect to θ is equivalent to the recently developed notion of effective bandwidth in communication networks when restricting to a family of two-state Markov modulated arrival processes. This equivalence relation has been recently extended to other Markov processes (see [29, 16]). Since our definition of MER with respect to θ is fairly general and does not require a preset mathematical model, our approach might be able to be used to obtain the effective bandwidth for other real-time traffic, e.g. video. Moreover, the tool for computing the bounds and approximations of the tail distributions of queues in a network is already available in our analysis once the MER with respect to θ of input processes are obtained. A tentative solution for admission control of high speed networks is proposed in [9]. Further numerical studies will be reported in a separate paper.

Acknowledgements:

The author would like to thank Roch Guérin [23], Jim Kurose [30], Khosrow Sohraby [37] and Pantelis Tsoucas [39] for their valuable discussions on their works. Insightful discussions at various stages with Leonidas Georgiadis, Armand Makowski, and Joy Thomas are gratefully acknowledged. Last, but not least, the author would like thank the referees for their careful reading and detailed comments that have greatly improved the presentation of this paper.

A Appendix A

In this appendix, we prove Lemma 3.6.

Proof. (i) Since $a(t) \leq M$, $A(s, s+t) \leq Mt < \infty$. Thus, $\|A(s, s+t)\|_\infty \leq Mt < \infty$. Since $\|A(s, s+t)\|_\infty$ is the greatest lower bound, for every $\epsilon > 0$ there exists a $\delta > 0$ such that $\text{Prob}(A(s, s+t) > \|A(s, s+t)\|_\infty - \epsilon) \geq \delta > 0$. From Chernoff's bound, it then follows that

$$\frac{1}{\theta} \log E e^{\theta A(s,s+t)} \geq \|A(s, s+t)\|_\infty - \epsilon + \frac{1}{\theta} \log \delta.$$

Letting $\theta \to \infty$ and then $\epsilon \to 0$ yields

$$\lim_{\theta \to \infty} \frac{1}{\theta} \log E e^{\theta A(s,s+t)} \geq \|A(s, s+t)\|_\infty.$$

Thus,

$$\lim_{\theta\to\infty} A^*(\theta,t) \geq \sup_{s\geq 0} \lim_{\theta\to\infty} \frac{1}{\theta} \log E e^{\theta A(s,s+t)} \geq \sup_{s\geq 0} \|A(s,s+t)\|_\infty.$$

(ii) Let $\phi(\theta) = E\exp(\theta X)$ be the the moment generating function of a bounded random variable X ($0 \leq X \leq M_1$ for some constant M_1). Then for any finite θ, the n^{th} derivative of $\phi(\theta)$, $n = 1, 2, \ldots$, denoted as $\phi^{(n)}(\theta)$, exists and equals to $E(X^n \exp(\theta X))$. In particular, the first derivative $\phi'(0)$ is equal to EX. Applying Taylor's expansion to the function $\log\phi(\theta)$ at $\theta = 0$ yields

$$\log\phi(\theta) = \theta EX + \frac{\theta^2}{2}\frac{\phi^{(2)}(\theta_1)\phi(\theta_1) - (\phi^{(1)}(\theta_1))^2}{(\phi(\theta_1))^2}$$

for some $\theta_1 \in [0,\theta]$. Since $0 \leq X \leq M_1$,

$$\frac{1}{\theta}\log\phi(\theta) \leq EX + \theta\frac{M_1^2}{2}. \tag{85}$$

Replacing X and M_1 with $A(s,s+t)$ and Mt in (85) and taking the supremum on both sides yields

$$A^*(\theta,t) \leq \sup_{s\geq 0} EA(s,s+t) + \theta\frac{(Mt)^2}{2}. \tag{86}$$

Letting $\theta \to 0$ completes the derivation for (ii).

(iii) Observe that $\log E\exp(\theta A(s,s+t))$ is bounded and convex in θ for $0 < \theta < \infty$ since $a(t) \leq M < \infty$. Since the supremum or upper limit of bounded and convex functions is still bounded and convex, both $\sup_{s\geq 0}\log E\exp(\theta A(s,s+t))$ and $\limsup_{t\to\infty}(1/t)\sup_{s\geq 0}\log E\exp(\theta A(s,s+t))$ are bounded and convex in θ for $0 < \theta < \infty$. It then follows from the boundedness and the convexity that $\theta A^*(\theta,t)$ and $\theta a^*(\theta)$ are continuous for $0 < \theta < \infty$. Multiplying by the continuous function $1/\theta$ completes the proof. $\square$

References

[1] D. Anick, D. Mitra and M.M. Sondhi, "Stochastic theory of a data handling system with multiple sources," *Bell Syst. Tech. J.*, Vol. 61, pp. 1871-1894, 1982.

[2] M. Avriel, "Nonlinear Programming: Analysis and Methods", *Prentice-Hall*, 1976.

[3] F. Baccelli and P. Bremaud. *Elements of Queueing Theory*. New York: Springer Verlag, 1990.

[4] R.E. Barlow and F. Proschan, *Stochastic theory of reliability and life testing.* Holt, Rinehart and Winston, Reading, M.A., 1975.

[5] A. Birman, H.R. Gail and S.L. Hantler, "An optimal service policy for buffer systems," *IBM RC 16641*, 1991.

[6] A.A. Borovkov, *Stochastic Processes in Queueing Theory.* New York: Springer-Verlag, 1976.

[7] A. Brandt, P. Franken and B. Lisek, *Stationary Stochastic Models.* New York:Wiley & Sons, 1990.

[8] J. Bucklew, *Large Deviation Techniques in Decision, Simulation and Estimation.* New York, NY: J. Wiley & Sons, Inc., 1990.

[9] C.S. Chang, "Stability, queue length and delay, Part II: Stochastic queueing networks," *IEEE Conference on Decision and Control*, Vol 1, pp. 1005-1010, 1992.

[10] C.S Chang, J.A. Thomas, and S.-H. Kiang, "On the stability of open networks: an unified approach by stochastic dominance," to appear in *Queueing Systems.*

[11] C.S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin, "Effective Bandwidth and Fast Simulation of ATM Intree Networks," *IBM RC 18586*, 1992.

[12] H. Chen and D.D. Yao, "Optimal scheduling control in a multi-class fluid network," *IEEE Conference on Decision and Control*, pp. 1105-1106, 1989.

[13] Y.S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales.* New York: Springer-Verlag, 1988.

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

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

[16] A.I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *preprint.*

[17] J.D. Esary, F. Proschan and D. Walkup, "Association of random variables, with applications," *Ann. Math. Statist.*, Vol. 38, pp. 1466-1474, 1967.

[18] S.G. Foss, "Some properties of open queueing networks," *Problems of Information Transmission*, Vol. 25, pp. 241-246, 1989.

[19] H.R. Gail, G. Grover, R. Guérin, S.L. Hantler, Z. Rosberg and M. Sidi, "Buffer size requirement under longest queue first," *IBM RC 16486*, 1991.

[20] R.J. Gibbens and P.J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Systems*, Vol. 9, pp. 17-28, 1991.

[21] S.J. Golestani, "Congestion-free transmission of real-time traffic in packet networks," *Proc. IEEE Infocom'90*, pp. 527-536, 1990.

[22] P.M. Gopal and B.K. Kadaba, "Network delay considerations for packetized voice," *Performance Evaluation*, Vol. 9, pp. 167-180, 1989.

[23] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.,* Vol. 9, pp. 968-981, 1991.

[24] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge, 1987.

[25] J.Y. Hui, "Resource allocation for broadband networks," *IEEE Select. Areas Commun.*, Vol. 6, pp. 1598-1608, 1988.

[26] F.P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, Vol. 9, pp. 5-16, 1991.

[27] J.F.C. Kingman, "Inequalities in the theory of queues," *J. Roy. Stat. Soc., Series B*, Vol. 32, pp. 102-110, 1970.

[28] J.F.C. Kingman, *Subadditive processes* in Ecole d'ete de probabilities de Saint-Flour, edited by A. Dold and B. Eckmann, Lecture Notes in Mathematics, 539, Springer-Verlag, Berlin, pp. 165-223, 1976.

[29] G. Kesidis, J. Walrand and C.S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *preprint.*

[30] J.F. Kurose, "On computing per-session performance bounds in high-speed multi-hop computer networks," *Proc. ACM SIGMETRICS and PERFORMANCE'92*, (Newport, Rhode Island, June 1992), pp. 128-139.

[31] R.M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proc. Camb. Phil. Soc.*, Vol. 58, pp. 497-520, 1962.

[32] S.H. Lu and P.R. Kumar, "Distributed scheduling based on due dates and buffer priorities," *IEEE Tran. Automat. Contr.*, Vol. 35, pp. 289-298, 1991.

[33] J.R. Perkins and P.R. Kumar, "Stable, distributed, real-time scheduling of flexible manufacturing/assembly/disassembly systems," *IEEE Tran. Automat. Contr.*, Vol. 34, pp. 139-148, 1989.

[34] M.J. Rider, "Protocols for ATM access networks," *Globecom'88.*

[35] S.M. Ross, *Stochastic Processes.* New York: J. Wiley & Sons, 1983.

[36] H.L. Royden, *Real Analysis.* New York: Macmillan, 1968.

[37] K. Sohraby, "On the asymptotic behavior of heterogeneous statistical multiplexer with applications," *INFOCOM'92*, Florence, Italy.

[38] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, Berlin: J. Wiley & Sons, 1983.

[39] P. Tsoucas and J. Walrand, "Monotonicity of throughput in non-Markovian networks, " *J. Appl. Prob.* 26, pp. 134-141, 1989.

[40] G.M. Woodruff, R.G.H. Rogers and P.S. Richards, "A congestion control framework for high-speed integrated packetized transport," *Globecom'88.*

[41] O. Yaron and M. Sidi, "Calculating performance bounds in communication networks," *IEEE INFOCOM'93*, Vol. 2, pp. 539-546.

[42] L. Zhang, "Designing a new architecture for packet switching communication networks," *IEEE Communications Magazine*, Vol. 25, No. 9, pp. 5-12, 1987.