On the Stability of Open Networks: an Unified Approach by Stochastic Dominance

Cheng-Shang Chang Joy A Thomas
IBM Research Division, T.J. Watson Research Center
P.O. Box 704, Yorktown Heights, NY 10598

Shaw-Hwa Kiang
Department of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027-6699

August 1992; revision April 1993

Abstract

Using stochastic dominance, in this paper we provide a new characterization of point processes. This characterization leads to a unified proof for various stability results of open Jackson networks where service times are i.i.d. with a general distribution, external interarrival times are i.i.d. with a general distribution and the routing is Bernoulli. We show that if the traffic condition is satisfied, i.e., the input rate is smaller than the service rate at each queue, then the queue length process (the number of customers at each queue) is tight. Under the traffic condition, the p^{th} moment of the queue length process is bounded for all t if the $p+1^{th}$ moment of the service times at all queues are finite. If, furthermore, the moment generating functions of the service times at all queues exist, then all the moments of the queue length process are bounded for all t. When the interarrival times are unbounded and non-lattice (resp. spread-out), the queue lengths and the remaining service times converge in distribution (resp. in total variation) to a steady state. Also, the moments converge if the corresponding moment conditions are satisfied.

Keywords: stability, queue, Jackson network, stochastic ordering

1 Introduction

Consider an open Jackson type of network with I single-server queues. Customers, arriving to the network according to a *stationary* renewal process with rate λ, are routed (independent of all else) to queue i with probability $p_{0,i}$, $i = 1, \ldots, I$. Assume that the buffer size at each queue is infinite and that the service discipline at each queue is nonpreemptive and work conserving, e.g., FCFS. The service times at queue i, $i = 1, \ldots, I$, are i.i.d. random variables with the same distribution as the random variable β_i. After completing service at queue i, a customer is routed (independent of all else) to queue j with probability $p_{i,j}$. So the routing mechanisms for both external and internal customers are Bernoulli. Let γ_i, $i = 1, \ldots, I$, be the solution of the usual traffic equations:

$$\gamma_i = \lambda_i + \sum_{j=1}^{I} \gamma_j p_{j,i}, \tag{1}$$

where $\lambda_i = \lambda p_{0,i}$ is the external arrival rate to queue i. Clearly, γ_i is the total arrival rate at queue i. Define $\rho_i = \gamma_i E\beta_i$. Let $q(t) = (q_1(t), \ldots, q_I(t))$ and $v(t) = (v_1(t), \ldots, v_I(t))$, where $q_i(t)$ is the number of customer at queue i (including the one being served) at time t, and $v_i(t)$ is the remaining service time (of the one being served) at time t. Also, let $u(t)$ be the remaining time to the next arrival.

Theorem 1.1 *Suppose the network starts from a fixed number of customer in each queue with a fixed remaining service time (if any).*

(i) *If $\rho_i < 1$ for all i, then the stochastic process $q(t)$ is tight, i.e., for every $\epsilon > 0$ there exists a finite vector $q = (q_1, \ldots, q_I)$ such that $\inf_t Prob(q(t) \leq q) \geq 1 - \epsilon$.*

(ii) *If $\rho_i < 1$ and $E\beta_i^{p+1}$ for all i and some $p > 0$, then $\sup_{t \geq 0} E \sum_{i=1}^{I} (q_i(t))^p < \infty$.*

(iii) *If $\rho_i < 1$ for all i and there exists a $\theta' > 0$ such that $E\exp(\theta' \beta_i) < \infty$ for all i, then there exists a $\theta > 0$ such that $\sup_{t \geq 0} E \exp(\theta \sum_{i=1}^{I} q_i(t)) < \infty$.*

Furthermore, suppose that the distribution $G(x)$ of external interarrival times is unbounded and non-lattice.

(i') *Under the assumptions in (i), $(q(t), v(t), u(t))$ converges in distribution to a random vector $(q(\infty), v(\infty), u(\infty))$, i.e., for any Borel set B*

$$\lim_{t \to \infty} Prob((q(t), v(t), u(t)) \in B) = Prob((q(\infty), v(\infty), u(\infty)) \in B). \tag{2}$$

The assumption of stationarity on the external arrival renewal process is not needed. Moreover, $(q(t), v(t), u(t))$ is ergodic, i.e.,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbf{1}\{(q(s), v(s), u(s)) \in B\} ds = Prob((q(\infty), v(\infty), u(\infty)) \in B) \ \ a.s. \tag{3}$$

If $G(x)$ is also spread out, then $(q(t), v(t), u(t))$ converges to $(q(\infty), v(\infty), u(\infty))$ in total variation, i.e., the convergence in (2) is uniform *for all Borel sets.*

(ii') *Under the assumptions in (ii),*

$$\lim_{t\to\infty} E(q_i(t))^p = E(q_i(\infty))^p < \infty, i = 1, \ldots, I. \tag{4}$$

(iii') *Under the assumptions in (iii), there exists a $\theta > 0$ such that*

$$\lim_{t\to\infty} E\exp(\theta q_i(t)) = E\exp(\theta q_i(\infty)) < \infty, i = 1, \ldots, I. \tag{5}$$

In Meyn and Tweedie [23, 24], the result in Theorem 1.1(i) is referred as the *topological stability*, while the result in Theorem 1.1(i') is classified as the *probabilistic stability.* For certain classes of Markov processes, these two notions of stability are shown to be equivalent (see Meyn and Tweedie [23, 24]). In this paper we establish the topoligical stability result in Theorem 1.1(i)(ii)(iii) using stochastic comparisons of point processes and queues. For completeness, we apply the classical theory of regenerative processes (e.g. Asmussen [1]) to obtain the probabilistic stability results, though it might be possible to apply the results in Meyn and Tweedie [23, 24].

For the topological stability, we note the external arrival process need not to be renewal. An extension of Theorem 1.1 to general arrival processes (satisfying certain regularity conditions developed in this paper, e.g., stationarity and ergodicity) is stated in Theorem 4.3. For a single queue, the results (i),(ii) and (iii) in Theorem 1.1 (or Theorem 4.3) are attributed to Loynes [21], Kiefer and Wolfowitz [16, 17] and Kingman [18], respectively. These stability results were established for a network of parallel queues in Rolski and Szekli [26]. Theorems 1.1 and 4.3 can be viewed as an extension of stability results from a single queue to an open network of queues. Part of the results in these two theorems are known in the literature. For instance, Sigman [30, 31] showed the result in (i') for two-queue networks under the renewal assumptions for both interarrival and service times. Using the equivalence result in [23, 24] of the topological stability and probability stability, Meyn and Down [22] obtained the first moment result ($p = 1$) in (ii'). Foss [12] showed that the result on the moment generating function in (iii) assuming that the service times and interarrival times satisfies the strong Cramer's condition in Borovkov [3]. Recently, Foss [13] (the translated version appeared while this paper was under review) obtained the result in (i') without the assumption of unbounded external interarrival times. The approaches used by Sigman [30, 31], Meyn and Down [22] and Foss [12, 13] are quite different and it is not clear how one obtains all the stability results in Theorem 1.1 from these approaches. While not claiming the originality of all these results for open networks, one of our objectives in this paper is to provide a simple and insightful approach that allows us to obtain all the stability results in Theorem 1.1. Our approach is similar to that in Meyn and Down [22] in the sense that we take advantage of the equivalence of the topological stability and probability stability. However, instead of using the first moment condition, we only need a weaker condition, stochastic dominance of point processes and queues, to establish the topological stability. To establish the probabilistic stability, we use an argument similar to that in Sigman [30] for regenerative processes. We note that the notion of stochastic dominance of queues was used in Foss [12, 13].

The paper is organized as follows. In §2, we stochastically bound point processes by introducing the notion of envelope processes. This leads a unified way to characterize a point process by two simple parameters: the rate λ and the (nonnegative) regularity variable σ. Intuitively, the regularity variable measures the deviation of a point process from a deterministic process with rate λ. In particular, when σ is upper bounded by a constant, this characterization reduces to the worst case bound in Cruz [8, 9], Kurose [20] and Chang [4]. When σ has a finite moment generating function, it recovers the characterization in Foss [12] and Chang [5]. A similar characterization to that in Meyn and Down [22] is when σ has a finite first moment. The advantage of this characterization, as demonstrated in Cruz [8, 9], Kurose [20], and Chang [5], is to provide an algebraic tool that can be used to obtain stochastic bounds for the queue length processes in a network, in terms of the regularity variables and the rates of arrival and service processes.

In §3, we consider a single queue with autonomous service in which the queue length process is completely characterized by an arrival process and a service process. Using the characterizations of the arrival process and the service process, we obtain a stochastic bound for the queue length process. The stochastic bound leads to various stability results for the queue length process. For instance, the p^{th} power of the queue length process is uniformly integrable if both the regularity variables of the arrival process and the service process have finite p^{th} moments. To link up with ordinary $G/GI/1$ queues, we consider $G/GI/1$ queues with vacations in which the vacations have the same distribution as the service times. We show that the difference between the queue length process of the queue with autonomous service and that of the vacation model is at most 1, when subject to the same arrival process. Moreover, the departure processes of both queues are thinning processes of renewal processes. This important property allows us to decompose a $G/GI/1$ queue with vacations into an arrival process and a service process. Also, it provides a stochastic bound for each queue in an open Jackson network with I queues by assuming the arrival process at each queue is a superposition of $I-1$ independent renewal processes and the external arrival process.

In §4, we prove Theorem 1.1 and Theorem 4.3. The argument consists of three steps: (1) Decompose the network into isolated queues, and stochastically bound each queue by assuming the departure processes from other queues are renewal with rates equal to service rates. (2) Stochastically bound the total number of customers in the network by the monotonicity results in Foss [12], Shanthikumar and Yao [29] and Chang [4]. (3) Apply the classical theory of regenerative processes to establish the connection between the topoligical stability and the probabilistic stability. To illustrate steps (1) and (2), consider a $GI/GI/1 \to /GI/1$ tandem queue. Assume that the arrival rate is λ and that the service rates at the first queue and the second queue are μ_1 and μ_2 respectively. The network is ergodic if $\lambda < \mu_1$ and $\lambda < \mu_2$. The argument goes as follows: The stability of the first queue follows from the result for a single queue. The arrival process of the second queue is the departure process from the first queue which is stochastically dominated by a renewal process with rate μ_1. By the monotonicity result in Foss [12] and Shanthikumar and Yao [29] , the second queue would have more customers if the arrival process were a process with rate μ_1. Thus, the second queue would be stable if $\mu_1 < \mu_2$. If $\mu_1 \geq \mu_2 > \lambda$, we consider an auxiliary tandem queue with $\mu_1' = (\mu_2 + \lambda)/2$. Now we have $\lambda < \mu_1' < \mu_2$ and the both queues of the auxiliary system are stable, as demonstrated by the pervious argument. Since we reduce the service rate from μ_1 to μ_1', we also expect the total number of customers in the original system to be smaller than that of the auxiliary system (see Foss [12]or Shanthikumar and Yao [29]). From the stability of the auxiliary system, we then prove the stability of the original system. The existence of such an auxiliary system in an open Jackson network is guaranteed in Chang [4].

2 Stochastic ordering and envelope processes

In this section, we will first briefly introduce the notion of stochastic ordering and use this notion to define a new characterization of point processes. A random variable X is stochastically smaller than Y, denoted by $X \leq_{st} Y$, if $Ef(X) \leq Ef(Y)$ for all f increasing. A stochastic process $\{X(t), t_1 \leq t \leq t_2\}$ is stochastically dominated by $\{Y(t), t_1 \leq t \leq t_2\}$, denoted by $\{X(t), t_1 \leq t \leq t_2\} \leq_{st} \{Y(t), t_1 \leq t \leq t_2\}$, if for all n and $t_1, \ldots, t_n \in [t_1, t_2]$

$$Ef(X(t_1), \ldots, X(t_n)) \leq Ef(Y(t_1), \ldots, Y(t_n)) \tag{6}$$

for all f increasing.

It is well known (e.g., Kamae, Krengel and O'Brien [15], Stoyan [32], Ross [28] and Rolski and Szekli [27] and references therein) that $X \leq_{st} Y$ if and only if $\text{Prob}(X \geq s) \leq \text{Prob}(Y \geq s)$ for all s. Moreover, $\{X(t), t_1 \leq t \leq t_2\} \leq_{st} \{Y(t), t_1 \leq t \leq t_2)\}$ if one can construct two stochastic processes $\{X'(t)\}$ and $\{Y'(t)\}$ on the same probability space such that $\{X'(t)\} =_{st} \{X(t)\}$, $\{Y'(t)\} =_{st} \{Y(t)\}$, and $X'(t) \le Y'(t)$ for all $t_1 \le t \le t_2$.

Now consider a point process A with $A(t_1, t_2)$ being the number of arrivals in the interval $(t_1, t_2]$, $-\infty < t_1 \le t_2 < \infty$.

Definition 2.1 *A stochastic process $\{\hat{A}(t), t \ge 0\}$ is an* upper envelope process *(UEP) of a point process A if for all $t \ge 0$*

$$\{A(t-s,t), 0 \le s \le t\} \le_{st} \{\hat{A}(s), 0 \le s \le t\}. \tag{7}$$

Similarly, a stochastic process $\{\breve{A}(t), t \ge 0\}$ is a lower envelope process *(LEP) of a point process A if for all $t \ge 0$*

$$\{A(t-s,t), 0 \le s \le t\} \ge_{st} \{\breve{A}(s), 0 \le s \le t\}. \tag{8}$$

In particular, if an UEP has the form, $\hat{A}(t) = at + \sigma$, for some constant a and some nonnegative random variable σ, then it will be called a linear UEP (a, σ). The notion of linear LEP's is defined similarly when an LEP has the form $\breve{A}(t) = at - \sigma$.

Definition 2.1 is a generalization of Cruz [8, 9], Kurose [20] and Chang [4, 5]. In these papers, the notion of envelope processes has been successfully applied to obtain performance bounds of various networks. To provide a unified view of these papers, we introduce the following characterization of point processes based on the notions of linear UEP's and LEP's.

Definition 2.2 *A point process A is said to be*

(R1a) *λ-tight from above (t.f.a.) if for all $\epsilon > 0$, there exists a linear UEP $(\lambda + \epsilon, \sigma_\epsilon^u)$ such that $\sigma_\epsilon^u < \infty$ a.s.,*

(R2a) *λ-L^p bounded from above (L^p b.f.a.) for some constant $p > 0$ if (R1a) holds and $E(\sigma_\epsilon^u)^p < \infty$,*

(R3a) *λ-moment generating function bounded from above (m.b.f.a.) if (R1a) holds and there exists a constant $\theta > 0$ such that $E\exp(\theta\sigma_\epsilon^u) < \infty$.*

(R4a) *λ-bounded from above (b.f.a.) if (R1a) holds and σ_ϵ^u is bounded above by a constant, i.e., $\sigma_\epsilon^u \le M$ for some constant $M < \infty$ (M may depend on ϵ),*

Similarly, we define the notions of λ-tight from below (t.f.b.), λ-bounded from below (b.f.b.), λ-L^p bounded from below (L^p b.f.b.) and λ-moment generating function bounded from below (m.b.f.b.) by replacing the linear UEP by the linear LEP $(\lambda - \epsilon, \sigma_\epsilon^l)$.

Clearly, b.f.a. $\Rightarrow$ m.b.f.a. $\Rightarrow$ L^p b.f.a. $\Rightarrow$ t.f.a. ($R4a \Rightarrow R3a \Rightarrow R2a \Rightarrow R1a$) and b.f.b. $\Rightarrow$ m.b.f.b. $\Rightarrow$ L^p b.f.b. $\Rightarrow$ t.f.b. If the random variable σ_ϵ^u in Definition 2.2 is a constant, then we obtain the characterization by Cruz [8, 9], Kurose [20] and Chang [4] for deterministic networks. If σ_ϵ^u has a finite moment generating function, then we recover the characterization by Foss [12] and Chang [5]. A similar characterization to that in Meyn and Down [22] can be obtained by assuming that σ_ϵ^u has a finite mean. We note that a similar concept to the characterization in Definition 2.2 was used in Rolski [25], pp. 123-124, to establish uniform integrability of virtual waiting times in queues with doubly stochastic Poisson arrivals.

Example 2.3 Consider a time stationary and ergodic point process A with rate λ. Clearly, its reversed time process, denoted by A^r, is also a time stationary and ergodic point process with rate λ. From Definition 2.1, the reversed time process started at time 0, $\{\bar{A}^r(t) \stackrel{\text{def}}{=} A^r(0,t), t \geq 0\}$, is both a UEP and an LEP of A. Moreover, we will argue that a stationary and ergodic point process with rate λ is both λ-t.f.a. and λ-t.f.b. Since the rate for the reversed time process is λ, $\lim_{t\to\infty} \bar{A}^r(t)/t = \lambda$ a.s. Thus, for every $\epsilon > 0$, we define two *finite* random variables as follows:

$$T^u = \sup\{t : \bar{A}^r(t) \geq (\lambda+\epsilon)t\}, \tag{9}$$

and

$$T^l = \sup\{t : \bar{A}^r(t) \leq (\lambda-\epsilon)t - 1\}. \tag{10}$$

Observe that $\bar{A}^r(t) \leq (\lambda+\epsilon)t$ for $t \geq T^u$ and that $\bar{A}^r(t) \leq \bar{A}^r(T^u) = (\lambda+\epsilon)T^u$ for $t < T^u$ since $\bar{A}^r(t)$ is increasing in t. Thus, $\bar{A}^r(t) \leq (\lambda+\epsilon)t + (\lambda+\epsilon)T^u$. This implies that $(\lambda+\epsilon, (\lambda+\epsilon)T^u)$ is a linear UEP of A. Similarly, observe that $\bar{A}^r(t) \geq (\lambda-\epsilon)t - 1$ for $t \geq T^l$ and that $\bar{A}^r(t) \geq 0$ for $t < T^l$. Thus, we have $\bar{A}^r(t) \geq (\lambda-\epsilon)t - 1 - (\lambda-\epsilon)T^l$. This implies that that $(\lambda-\epsilon, (\lambda-\epsilon)T^l + 1)$ is a linear LEP of A.

Example 2.4 Consider a (time) stationary renewal point process A. Assume that the inter-arrival times are identically distributed as a random variable β with $E\beta = 1/\lambda < \infty$. Thus, the stationary renewal point process has rate λ. Moreover, the reversed time process A^r is reversible, i.e., it has the same joint distribution as A. This implies that the reversed time process started at time 0, $\bar{A}^r(t)$, is the delayed renewal process with the first point having the equilibrium distribution of β. Since a stationary renewal process is stationary and ergodic (Daley and Vere-Jones [11], Exercise 10.3.1), it follows from Example 2.3 that for every $\epsilon > 0$, there exists two nonnegative *finite* random variables $\sigma_\epsilon^u = (\lambda+\epsilon)T^u$ in (9) and $\sigma_\epsilon^l = (\lambda-\epsilon)T^l - 1$ in (10) such that $(\lambda+\epsilon, \sigma_\epsilon^u)$ is a linear UEP of A and $(\lambda-\epsilon, \sigma_\epsilon^l)$ is a linear LEP of A.

By establishing some properties of the two random variables σ_ϵ^u and σ_ϵ^l in Example 2.4, we have the following results for a stationary renewal process. The proof is given in Appendix A.

Lemma 2.5 *For a stationary renewal process with the generic interarrival time β ($E\beta = 1/\lambda$), it is*

(i) λ-*m.b.f.a.,*

(ii) λ-*m.b.f.b. if there exists a* $\theta' > 0$ *such that* $E\exp(\theta'\beta) < \infty$,

(iii) λ-L^p *b.f.b. if* $E\beta^{p+1} < \infty$,

(iv) λ-*t.f.b.*

Use the closure property of stochastic ordering, one can easily verify the following lemma for the superposition of independent point processes.

Lemma 2.6 *Let A_i, $i = 1, \dots, m$, be m* independent *point processes. If A_i has a linear UEP (resp. LEP) (a_i, σ_i), then the superposition process $A = \sum_{i=1}^m A_i$ has a linear UEP (resp. LEP) $(\sum_{i=1}^m a_i, \sum_{i=1}^m \sigma_i)$. Hence, the superposition process A is λ-t.f.a. (resp. t.f.b.) if each point process A_i is λ_i-t.f.a. (resp. t.f.b.), where $\lambda = \sum_{i=1}^m \lambda_i$. Similar results hold for other characterizations in (R2-4).*

3 Uniform integrability, tightness and queues with autonomous service

In this section, we will first review the notions of uniform integrability and tightness. We then establish connections between these notions with the characterization of point processes introduced in the previous section by considering queues with autonomous service.

Consider a stochastic process $\{X(t), t \geq 0\}$. Let $\{F_t(x) = \text{Prob}(X(t) < x), t \geq 0\}$ be the family of distribution functions associated with the stochastic process $X(t)$. A real, continuous function g is called *uniformly integrable* (u.i.) relative to $\{F_t\}$ if

$$\lim_{a\to\infty} \sup_{t\geq 0} \int_{|x|\geq a} |g(x)| dF_t(x) = 0. \tag{11}$$

In particular, the process $X(t)$ is *tight* (or bounded in probability) if the function 1 is u.i. relative to $\{F_t\}$ (see Chow and Teicher [6]). This is equivalent to

$$\lim_{a\to\infty} \inf_{t\geq 0} \text{Prob}(|X(t)| < a) = 1. \tag{12}$$

Intuitively, this implies that the process $X(t)$ does not blow up for any value of t.

For a function g that satisfies $\lim_{|x|\to\infty}|g(x)| = \infty$, it is known (Chow and Teicher [6], Proposition 8.1.2) that g is u.i. relative to $\{F_t\}$ if and only if the stochastic process $\{g(X(t)), t \geq 0\}$ is u.i., i.e.,

$$\lim_{a\to\infty} \sup_{t\geq 0} E\mathbf{1}_{\{|g(X(t))|\geq a\}}|g(X(t))| = 0. \tag{13}$$

An important property of uniform integrability is that if the process converges, the corresponding expectation also converges. The precise statement in given in the following proposition (Chow and Teicher [6], Theorem 8.1.2).

Proposition 3.1 *Suppose that $X(t)$ converges in distribution to a finite random variable $X(\infty)$ and that g is a nonnegative, continuous function for which $\sup_{t\geq 0}\int_{-\infty}^{\infty} g(x)dF_t(x) < \infty$. Then*

$$\lim_{t\to\infty} \int_{-\infty}^{\infty} g(x)dF_t(x) = \int_{-\infty}^{\infty} g(x)dF_\infty(x) < \infty \tag{14}$$

if and only if g is u.i. relative to $\{F_t\}$.

For a function g that satisfies $\lim_{|x|\to\infty}|g(x)| = \infty$, the convergence in (14) holds if $\{g(X(t)), t \geq 0\}$ is u.i. This is due to the fact that the uniform integrability of $\{g(X(t)), t \geq 0\}$ implies that $\sup_{t\geq 0} Eg(X(t)) < \infty$ (see Chow and Teicher [6], Theorem 4.2.1). In particular, we are interested in the two continuous functions $g(x) = |x|^p$ and $g(x) = \exp(\theta|x|)$, $\theta > 0$. Note that these two functions satisfy $\lim_{|x|\to\infty}|g(x)| = \infty$. Since we are interested in the stability of queueing networks, we will focus only on *nonnegative* processes and the sign of absolute value will be removed throughout.

Now we derive connections between tightness, uniform integrability and stochastic ordering. The proof is direct.

Lemma 3.2 *Consider two* nonnegative *stochastic processes $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$.*

(i) *If $X(t) \leq_{st} Y(t)$ for all t and $\{Y(t), t \geq 0\}$ is tight, then $\{X(t), t \geq 0\}$ is tight. In particular, if there exists a* finite *random variable Z such that $X(t) \leq_{st} Z$ for all t, then $\{X(t), t \geq 0\}$ is tight.*

(ii) *If $X(t) \leq_{st} Y(t)$ for all t and $\{g(Y(t)), t \geq 0\}$ is u.i. for some g increasing, then $\{g(X(t)), t \geq 0\}$ is u.i. In particular, if $X(t) \leq_{st} Z$ for all t and $Eg(Z) < \infty$ for some g increasing, then $\{g(X(t)), t \geq 0\}$ is u.i.*

Based on (12), one could define the tightness of a *nonnegative* vector process $\mathbf{X}(t) = (X_1(t), \ldots, X_m(t))$, i.e., for every $\epsilon > 0$, there exists a vector $(x_1, \ldots, x_m)$ such that

$$\inf_{t \geq 0} \text{Prob}(X_i(t) < x_i, i = 1, \ldots, m) \geq 1 - \epsilon. \tag{15}$$

The following proposition is well known.

Proposition 3.3 *A nonnegative vector process* $\mathbf{X}(t)$ *is tight if and only if each marginal process* $X_i(t)$, $i = 1, \ldots, m$, *is tight.*

Proposition 3.3 implies that the tightness of a queueing network might be obtainable by analyzing each queue in isolation, which is the approach we will use in the next section.

As a direct consequence of Proposition 3.3, the nonnegative process $\{\sum_{i=1}^m X_i(t), t \geq 0\}$ is tight if and only if each marginal process is tight. As we mentioned earlier, tightness is a notion of u.i. relative to a family of distributions when the function g is 1. In the paper, we are interested in the functions $g(x) = x^p$ (the p^{th} moment) and $g(x) = \exp(\theta x)$ (the moment generating function) for some $\theta > 0$. The following lemma shows the closure property of u.i. under summation for these two functions.

Lemma 3.4 *Consider* m *nonnegative processes* $X_i(t), i = 1, \ldots, m$.

(i) *If* $\{X_i(t), t \geq 0\}$, $i = 1, \ldots, m$, *are tight, then* $\{\sum_{i=1}^m X_i(t), t \geq 0\}$ *is tight.*

(ii) *If* $\{(X_i(t))^p, t \geq 0\}$, $i = 1, \ldots, m$, *are u.i. for some* $p > 0$, *then* $\{(\sum_{i=1}^m X_i(t))^p, t \geq 0\}$ *is also u.i.*

(iii) *If* $\{\exp(\theta' X_i(t)), t \geq 0\}$, $i = 1, \ldots, m$, *are u.i. for some* $\theta' > 0$, *then* $\{\exp(\theta \sum_{i=1}^m X_i(t)), t \geq 0\}$ *is also u.i., where* $\theta = \theta'/m$

The proof of (i) follows from Proposition 3.3. To prove (ii) and (iii), use the inequalities $\sum_{i=1}^m x_i \leq m \max_{1 \leq i \leq m} x_i$, and for $x_i \geq 0$, $i = 1, \ldots, m$, $\max_{1 \leq i \leq m} x_i \leq \sum_{i=1}^m x_i$.

Now consider a single queue with autonomous (gated) service. The queue length process of such a queue is characterized by two point processes: the arrival process A and the service process S. The queue is operated as follows: the queue length is increased by 1 whenever there is an arrival from the arrival process, and the queue length is decreased by 1 whenever there is an arrival from the service process and the queue is not empty. Let $q^a(t)$ denote the number of customers in the queue at time t. Then the queue length is governed by the following equation (Borovkov [2] (8.3)):

$$q^a(t) = \sup_{0 \le s \le t} (q^a(0) + A(0,t) - S(0,t), A(t-s,t) - S(t-s,t)). \tag{16}$$

Lemma 3.5 *If (i) the arrival process A has a linear UEP (a, σ), (ii) the service process S has a linear LEP (b, δ), and (iii) $a \le b$, then there exists a vector process $(X_1(t), X_2(t))$ such that for all $t \ge 0$, $q^a(t) \le_{st} q^a(0) + X_1(t) + X_2(t)$, $X_1(t) \le_{st} \sigma$ and $X_2(t) \le_{st} \delta$.*

Proof. From (16), it follows that

$$q^a(t) \le \sup_{0 \le s \le t} (A(t-s,t) - S(t-s,t)) + q^a(0). \tag{17}$$

Let $c = (b-a)/2 \ge 0$. It is easy to see that

$$\sup_{0 \le s \le t} (A(t-s,t) - S(t-s,t)) \le X_1(t) + X_2(t),$$

where $X_1(t) = \sup_{0 \le s \le t} (A(t-s,t) - (a+c)s)$ and $X_2(t) = \sup_{0 \le s \le t} ((b-c)s - A_2(t-s,t))$. From Definition 2.1, $X_1(t) \le_{st} \sigma$ and $X_2(t) \le_{st} \delta$. $\square$

We note that the arrival process A and the service process S need not to be mutually independent. The decomposition approach in Lemma 3.5 is also reported in a recent paper by Daley and Rolski [10], where they impose some mixing conditions on the interarrival times to show finiteness of waiting-time moments of $G/GI/1$ queues. As a direct application of Lemma 3.5, Lemma 3.2 and Definition 2.2, we have the following corollary for the queue length process of a queue with autonomous service.

Corollary 3.6 *Consider an autonomous service queue with an arrival process A and a service process S.*

(i) *If (1) A is λ-t.f.a., (2) S is μ-t.f.b., (3) $\lambda < \mu$, and (4) $q^a(0)$ is a* finite *random variable, then $\{q^a(t), t \ge 0\}$ is tight.*

(ii) *If (1) A is λ-L^p b.f.a., (2) S is μ-L^p b.f.b., (3) $\lambda < \mu$, and (4) $E(q^a(0))^p < \infty$, then $\{(q^a(t))^p, t \ge 0\}$ is u.i.*

(iii) *If (1) A is λ-m.b.f.a., (2) S is μ-m.b.f.b., (3) $\lambda < \mu$, and (4) $E\exp(\theta' q^a(0)) < \infty$ for some $\theta' > 0$, then there exists a $\theta > 0$ such that $\{\exp(\theta q^a(t)), t \ge 0\}$ is u.i.*

(iv) *If (1) A is λ-b.f.a., (2) S is μ-b.f.b., (3) $\lambda < \mu$, and (4) $q^a(0)$ is bounded, then* $\sup_{t \geq 0} q^a(t)$ *is bounded.*

We note that it is possible to show the converse statements in Corollary 3.6 for *stationary* point processes. Consider the queue length process, $q^a(t)$, of a single queue with autonomous service. Suppose $q^a(0) = 0$ and the service process is deterministic with rate μ. By Corollary 3.6 (i), $\{q^a(t), t \geq 0\}$ is tight if the arrival process A is λ-t.f.a. and $\lambda < \mu$. Conversely, if there exists a λ such that for all $\mu > \lambda$, $\{q^a(t), t \geq 0\}$ is tight then the stationary process A is λ-t.f.a.. The proof is analogous to that in Appendix A and will not be given here. Similar results hold for the other regularity notions. Intuitively, one could view feeding a stationary point process of rate λ into an autonomous queue with deterministic service process with rate μ. The queue length process is then bounded by the regularity variable, and this can be taken as the operational definition of the regularity notion t.f.a.

To link up with a ordinary $G/GI/1$ queue, we consider a G/GI/1 queue with vacations that have the same distributions as the service times. Recall that in a standard vacation model the server takes a vacation when the queue becomes empty. After the vacation is over, the server takes another vacation if the queue is still empty. Otherwise, it starts serving the customers in the queue. Now we compare the queue length process of such a vacation model with the queue length process of the autonomous service model. We assume that both queues have the same arrival process and that the service process of the queue with autonomous service is a stationary renewal process. Also, the interarrival times of the service process, the service times and vacations in the queue with vacations are i.i.d. Let $q^v(t)$ and $q^a(t)$ denote the queue length processes of the former (vacation model) and the latter.

Lemma 3.7 *Suppose that both queues start from the same number of customers at time 0 and that the remaining service time of the customer being served (if any) in the vacation model has the equilibrium distribution of the service times, then $q^a(t) \leq_{st} q^v(t) \leq_{st} q^a(t) + 1$ for all t.*

The proof of Lemma 3.7 is given in Appendix B. As in a queue with autonomous service, an important observation of such a vacation model is that *its departure process is a thinning process of a stationary renewal process with the same mean as the service times* (for details, see Appendix B). In the next section, we will use this to obtain a stochastic bound for an open network by analyzing each queue in isolation.

4 Stability of open networks of queues

In this section, we will prove the stability results in Theorem 1.1. First, we introduce a monotonicity result for an open network that we will use in the proof. Consider an open network $\mathcal{N}$

with I single-server FCFS queues. Denote by $\tau_i(n)$ the n^{th} external arrival epoch of queue i. Let $\beta_i(n)$ be the service time of the n^{th} customers in queue i (not necessary the n^{th} external arrival) and $r_i(n)$ be the index of the queue that the n^{th} customers in queue i is routed to after its service ($r_i(n) = 0$ if the customer leaves the network). Assume that the network is empty at time 0. It is clear that the behavior of the network $\mathcal{N}$ in the time interval $[0, \infty)$ is completely determined by the sequences $(\tau_i(n), \beta_i(n), r_i(n)), i = 1, \ldots, I, n = 1, 2, \ldots$. Denote by $D_i(n)$ the departure time of the n^{th} customers from queue i. The following proposition (e.g. see Foss [12], Shanthikumar and Yao [29]) states that the departure times are increasing in the arrival times and the services times. (A similar result for closed networks was shown by Tsoucas and Walrand [33].)

Proposition 4.1 *Consider two queueing networks:*

$$\mathcal{N} = \{(\tau_i(n), \beta_i(n), r_i(n)\} \ \ \textit{and} \ \ \mathcal{N}' = \{(\tau_i'(n), \beta_i'(n), r_i'(n)\}.$$

If for all $n \geq 1$ and $i = 1, \ldots, I$, $\tau_i(n) \leq \tau_i'(n)$, $\beta_i(n) \leq \beta_i'(n)$, $r_i(n) = r_i'(n)$ and $\sum_{n=1}^{\infty} \beta_i(n) = \infty$, then $D_i(n) \leq D_i'(n)$ for all i and n. (The inequality holds in the sample path sense.)

Now consider a special case of the network described above. We assume that $\{\beta_i, \beta_i(n), n = 1, 2, \ldots\}, i = 1, \ldots, I$, are sequences of i.i.d. random variables with $E\beta_i = 1/\mu_i$, and independent of all else. Thus, the assumption $\sum_{n=1}^{\infty} \beta_i(n) = \infty$ in Proposition 4.1 is satisfied almost surely for each sample path. Also assume that $\{r_i, r_i(n), n = 1, 2, \ldots\}, i = 1, \ldots, I$, are sequences of i.i.d. random variables with $\text{Prob}(r_i = j) = p_{i,j}, j = 0, 1, 2, \ldots, I$, and independent of all else. Moreover, the I arrival processes are also independent of all else, but they may depend on each other. We will call a network that satisfies these assumptions a generalized Jackson's network. We will denote such a network by $\mathcal{N}^J$ and the queue length process in queue i by $q_i(t)$, $i = 1, \ldots, I$.

Without loss of generality, we may further assume that the service policy at each queue is only nonpreemptive and work-conserving, and that there is no immediate feedback, i.e., $p_{i,i} = 0$ for all i. Since the service times are i.i.d., the number of customers in each queue is stochastically the same among the class of nonpreemptive and work-conserving policies. This class of polices include both the nonpreemptive FCFS policy and the nonpreemptive LCFS policy. Since the customers being feeding back immediately are served immediately under the nonpreemptive LCFS policy, we may adjust the routing probabilities such that $p_{i,i} = 0$ for all i.

Lemma 4.2 *Consider a generalized Jackson's network $\mathcal{N}^J$. Suppose the network starts from (fixed) $q_i(0)$, $i = 1, \ldots, I$ customers in queue i with (fixed) v_i being the remaining service* *time of the customer being served at queue i (if any). Let $v_{\max} = \max_i(v_i)$ and $A_i(0,t), i = 1,\ldots,I$ be the number of external arrivals to queue i in $(0,t]$. Then there exists a vector process $(q_1^v(t),\ldots,q_I^v(t))$ such that for all $t \geq 0$,*

(i) $\sum_{i=1}^I q_i(t) \leq_{st} \sum_{i=1}^I q_i^v(t)$,

(ii) $q_i^v(t) \leq_{st} q_i^a((t-v_{\max})^+) + 1$ *for all i,*

where $x^+ = \max(0,x)$ and $q_i^a(t)$ is the queue length process of a queue with autonomous service in which (1) $q_i^a(0) = q_i(0) + A_i(0,v_{\max}) + 1$, (2) the service process is a stationary renewal process with interarrival times being distributed as β_i, and (3) the arrival process is a superposition of $I-1$ stationary renewal processes with mean $1/(\mu_j p_{j,i})$, $j \neq i$, and the external arrival process to queue i.

Proof. Consider an auxiliary network $\mathcal{N}^v$ that coincides $\mathcal{N}^J$ except the remaining service times v_i and the service policy. In $\mathcal{N}^v$, the remaining service times are $v_{\max} + \beta_i^e, i = 1,\ldots,I$, where β_i^e has the equilibrium distribution of β_i, i.e., $\text{Prob}(\beta_i^e \geq x) = \int_x^\infty \text{Prob}(\beta_i \geq y)/E\beta_i dy$. Also, the server at every queue of $\mathcal{N}^v$ takes a vacation (when the queue is empty) with the vacation period being independent (of all else) and identically distributed as the service times at that queue. Denote by $q_i^v(t)$ the queue length process of queue i in $\mathcal{N}^v$. Note that the remaining service times are increased in $\mathcal{N}^v$ and that the server on vacation also increases the service times of customers that enter an empty queue. It then follows from Proposition 4.1 that the number of departures from each queue by time t in $\mathcal{N}^v$ is not greater than that in $\mathcal{N}^J$ for each sample path. Since the external arrival processes for both networks are identical, as a direct consequence, the total number of customers at time t in $\mathcal{N}^v$ is not less than that of $\mathcal{N}^J$ for each sample path. Thus, we have $\sum_{i=1}^I q_i(t) \leq_{st} \sum_{i=1}^I q_i^v(t)$ for all t.

Up to time $v_{\max}$, there are no departures from queues in $\mathcal{N}^v$. Thus, $q_i^v(t) = q_i(0) + A_i(0,t) \leq q_i(0) + A_i(0,v_{\max})$ for $0 \leq t < v_{max}$. As we discussed in the previous section and Appendix B, after time $v_{\max}$ the departure process from queue i in $\mathcal{N}^v$ is a thinning process of a stationary renewal process with rate μ_i. Since we assume that the routing is Bernoulli and independent of all else, the process formed by the customers that are routed from queue i to queue j is a thinning process of a stationary renewal process with rate $\mu_i p_{i,j}$. Thus, the arrival process at queue j is (stochastically) upper bounded by a superposition of $I-1$ independent stationary renewal processes with rate $\mu_i p_{i,j}$ and the external arrival process to queue j. From the monotonicity result in Proposition 4.1 (for a single queue), it follows that $q_j^v(t)$ is bounded above by the queue length of an autonomous queue subject to the upper bound arrival process. Applying Lemma 3.7 (for the upper bound arrival process) completes the proof. □

Theorem 4.3 *Consider a generalized Jackson's network $\mathcal{N}^J$ as described above. Suppose that the external arrival process to queue i, $i = 1, \dots, I$, denoted as A_i, is λ_i-t.f.a., and that the network starts from (fixed) $q_i(0)$, $i = 1, \dots, I$ customers in queue i with (fixed) v_i being the remaining service time of the customer being served at queue i (if any). Let γ_i's be the unique solution of*

$$\lambda_j + \sum_{i \neq j} \gamma_i p_{i,j} = \gamma_j, j = 1, \dots, I.$$

Then under the traffic condition $\gamma_i < \mu_i$ for all i,

(i) $\{\sum_{i=1}^I q_i(t), t \geq 0\}$ *is tight,*

(ii) $\{\sum_{i=1}^I q_i(t))^p, t \geq 0\}$ *is u.i. if A_i is λ_i-L^p b.f.a. and $E(\beta_i)^{p+1} < \infty$ for all i,*

(iii) $\{\exp(\theta \sum_{i=1}^I q_i(t)), t \geq 0\}$ *is u.i. for some $\theta > 0$ if A_i is λ_i-m.b.f.a. and $E \exp(\theta' \beta_i) < \infty$ for some $\theta' > 0$ and for all i.*

Proof. It is known (e.g. Chang [4] §5 (P2), Foss [13], Lemma 3) that for every vector $(\mu_1, \dots, \mu_I)$ satisfying $\gamma_j < \mu_j, j = 1, \dots, I$, there exists a vector $(\mu_1^s, \dots, \mu_I^s)$ such that $\mu_j^s \leq \mu_j$ for all j, and

$$\lambda_j + \sum_{i \neq j} \mu_i^s p_{i,j} < \mu_j^s, \quad j = 1, \dots, I. \tag{18}$$

Thus, there is an auxiliary network $\mathcal{N}^s$ that coincides $\mathcal{N}^J$ except the services times at queue i are scaled by a factor $\mu_i/\mu_i^s \geq 1$ for all i. It then follows from Proposition 4.1 that $\sum_{i=1}^I q_i(t) \leq_{st} \sum_{i=1}^I q_i^s(t)$ for all t, where $q_i^s(t)$ is the queue length process of queue i in $\mathcal{N}^s$. By Lemma 3.2, it suffices to prove the stability results for the network $\mathcal{N}^s$.

From Lemma 4.2 and the closure properties under summation in Lemma 3.4, we only need to establish the stability results for $q_j^a(t)$, $j = 1, \dots, I$, with $q_j^a(0) = q_j(0) + A_j(0, v_{\max}) + 1$, where $v_{\max} = \max_i(v_i)$. The arrival process of $q_j^a(t)$ is a superposition of $I - 1$ *independent* stationary renewal processes with mean $1/(\mu_i^s p_{i,j})$, $i \neq j$, and the external arrival process to queue j. By the assumption that the external arrival process to queue j is λ_j-t.f.a. (resp. L^p b.f.a., m.b.f.a.) and Lemmas 2.5(i) and 2.6, the arrival process of $q_j^a(t)$ is $\lambda_j + \sum_{i \neq j} \mu_i^s p_{i,j}$-t.f.a. (resp. L^p b.f.a., m.b.f.a.). Since the service process of $q_j^a(t)$ is simply a stationary renewal process with mean $1/\mu_j^s$, it then follows from Lemma 2.5(iv) (resp. Lemma 2.5(ii) and $E(\beta_j)^{p+1} < \infty$, Lemma 2.5(iii) and $E \exp(\theta' \beta_j) < \infty$ for some $\theta' > 0$) that the service process of $q_j^a(t)$ is μ_j^s-t.f.b. (resp. L^p b.f.b., m.b.f.b.). By Definition 2.2, this implies that $A_j(0, v_{\max})$, $j = 1, \dots, I$, is a finite random variable (resp. $EA_j(0, v_{\max})^p < \infty$, $E\exp(\theta'' A_j(0, v_{\max})) < \infty$ for some $\theta'' > 0$). Since (18) is satisfied, applying Corollary 3.6 completes the proof. $\square$

Proof.(Theorem 1.1) Using the properties for a stationary renewal process in Lemma 2.5, it then follows from Theorem 4.3 that (i)(ii) and (iii) holds (tightness of the joint queue length process follows from Proposition 3.3).

To show that (i'), let $\psi(t) = (q(t), v(t), u(t))$, where $q(t)$ is the queue length process, $v(t) = (v_1(t), \ldots, v_I(t))$ with $v_i(t)$ representing the remaining service time at queue i at time t, and $u(t)$ is the remaining time of the next external arrival. Since we assume that the distribution of external interarrival times are unbounded, $\psi(t)$ is regenerative with the obvious regenerative points when external arrivals see an empty network. (Arrival epochs finding an empty network form a renewal process.) If the distribution of the cycle length is non-lattice (resp. spread out) with finite mean, then it follows from the theory for the regenerative processes (e.g., Asmussen [1], Theorem V.1.2 for non-lattice cycles, Corollary VI.1.4 for spread out cycles, and Theorem V.3.1 for ergodicity) that (i') holds for any finite initial conditions. (Thus, the stationarity of the external arrival process is not needed). Since we assume that the distribution of external interarrival times is non-lattice (resp. spread out), the distribution of the cycle length (a random sum of external interarrival times) is non-lattice (resp. spread out) if the cycle length has a finite mean (Asmussen [1], Proposition VIII.3.2). Thus, we only need to show that the cycle length has a finite mean. This is equivalent to that $\psi(t)$ is positive recurrent.

Our approach is similar to that in Sigman [30]. We show that $\psi(t)$ is recurrent first, i.e., the network will be empty infinitely often. Consider the auxiliary network $\mathcal{N}^v$ in the proof of Lemma 4.2. Using the construction in Appendix B, the Markov process associated with $\mathcal{N}^v$ can be represented as $\psi^v(t) = (q^v(t), v^v(t), \zeta(t), u(t))$, where $q^v(t) = (q_1^v(t), \ldots, q_I^v(t))$ is the queue length process, $v^v(t) = (v_1^v(t), \ldots, v_I^v(t))$ denotes the remaining times to the next renewals of the service processes, $\zeta(t) = (\zeta_1(t), \ldots, \zeta_I(t))$ denotes vacation flags, and $u(t)$ is the remaining time to the next external arrival. From the monotonicity result in Lemma 4.2, it follows that the original network is empty if the auxiliary network is empty. Consider the events (indexed by x) $\mathcal{E}_x = \{\sum_{i=1}^I q_i^v(t) + v_i^v(t) \le x\}$, $x \ge 0$. It suffice to show that

$$\limsup_{t\to\infty} \mathbf{1}\{\mathcal{E}_0\} = 1 \quad a.s. \tag{19}$$

Before we proceed the proof for (19), we show that $\psi^v(t)$ is tight. Following the same argument as in the proof of Theorem 4.3, the process $\{\sum_{i=1}^I q_i^v(t), t \ge 0\}$ is tight. Since $v_i^v(t)$ is the the remaining time of a renewal process, it converges to the equilibrium distribution of β_i. Hence, $\{v_i^v(t), t \ge 0\}$ is also tight for all i. Similarly, $\{u(t), t \ge 0\}$ is tight. That $\{\zeta_i(t), t \ge 0\}$ is tight is trivial since it only takes values 0 and 1.

Since $q^v(t)$ and $v^v(t)$ are tight, there is a $x > 0$ (sufficiently large) such that $\limsup_{t\to\infty} \mathbf{1}\{\mathcal{E}_x\} = 1$ a.s. Note that if the event $\mathcal{E}_x$ occurs at an external arrival, then the event $\mathcal{E}_0$ will occur at the next external arrival with positive probability since we assume that the external interarrival times are unbounded. Using a standard geometrical trial argument, one concludes that (19) holds. Thus, $\psi^v(t)$ is recurrent. Let $\tau_n^v, n \geq 1$ be the epochs that arrivals see an empty network in $\mathcal{N}^v$. It then follows from the time average property of a recurrent regenerative process (e.g., Asmussen [1], Theorem V.3.1) and the bounded convergence theorem that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \mathrm{Prob}(\sum_{i=1}^{I} q_i^v(s) + v_i^v(s) \leq x) ds = \frac{E \int_{\tau_1^v}^{\tau_2^v} \mathbf{1}\{\sum_{i=1}^{I} q_i^v(t) + v_i^v(t) \leq x\} dt}{E(\tau_2^v - \tau_1^v)} \quad a.s. \tag{20}$$

We first note that the numerator in (20) is finite. If it is not, then by an geometric trial argument there would be an infinite number of regenerations in a cycle and we reach a contradiction. From tightness, the left hand side of (20) is positive if x is sufficiently large. Thus, $E(\tau_2^v - \tau_1^v) < \infty$ and $\psi^v(t)$ is positive recurrent. As an immediate result of stochastic dominance, $\psi(t)$ is also positive recurrent.

That (ii) and (iii) hold when the external arrivals form a *stationary* renewal process follows from the property of uniform integrability in Proposition 3.1. (Note that the stationarity of the external arrival process cannot be removed since we are not able to bound the difference before the external arrival process couples with its stationary one). □

Acknowledgements:

We are greatly indebted to Karl Sigman for valuable discussions regarding the proof of Theorem 1.1(1') and his paper [30]. We are also grateful to Tomasz Rolski for pointing out the connection of our decomposition approach to his works [25, 10]. Last, but not least, we 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 2.5. First, we construct the reversed time process, $\bar{A}^r(t)$, of the stationary renewal process as follows: construct a sequence of i.i.d. random variables $\{\beta, \beta(n), n = 2, \ldots\}$. Let $\beta(1)$ be a random variable (independent of $\beta(n)$'s) with the distribution $\mathrm{Prob}(\beta(1) \geq x) = \lambda \int_x^\infty \mathrm{Prob}(\beta \geq y) dy$. Let $\sum_{i=1}^n \beta(n)$ be the n^{th} arrival epoch of $\bar{A}^r(t)$, denoted by $\tau(n)$. Note that $\bar{A}^r(\tau(n)) = n$. Since σ_ϵ^u and σ_ϵ^l are linear in T^u and T^l, it suffices to show the following three properties for T^u and T^l (Note that it follows from Example 2.3 that both T^u and T^l are finite).

(P1) There exists a $\theta > 0$ such that $E\exp(\theta T^u) < \infty$.

(P2) If there exists a $\theta' > 0$ such that $E\exp(\theta'\beta) < \infty$, then there exists a $\theta > 0$ such that $E\exp(\theta T^l) < \infty$.

(P3) If $E\beta^{p+1} < \infty$ for some $p > 0$, then $E(T^l)^p < \infty$.

(P1) Let $\tilde{T}^u = \sup\{n : n \geq (\lambda+\epsilon)\tau(n)\}$ ($\tilde{T}^u = 0$ if $n < (\lambda+\epsilon)\tau(n)$ for all n). One observes that $\bar{A}^r(T^u) = \tilde{T}^u = (\lambda+\epsilon)T^u$. Thus, it suffices to show that there exists some $\theta > 0$ such that $E\exp(\theta\tilde{T}^u) < \infty$. First, we derive a stochastic ordering result for $\tilde{T}^u$. Let

$$\epsilon_1 = \frac{\epsilon}{2\lambda(\lambda+\epsilon)} \quad \text{and} \quad \tilde{\beta}(i) = \frac{1}{\epsilon_1}\Big(\frac{1}{\lambda} - \epsilon_1 - \beta(i)\Big).$$

Recall that $\tau(n) = \sum_{i=1}^n \beta(i)$ in our construction. Thus,

$$\begin{aligned}
&\mathrm{Prob}(\tilde{T}^u \geq s) = \mathrm{Prob}(n \geq (\lambda+\epsilon)\tau(n), \text{ for some } n \geq s)\\
&= \mathrm{Prob}(\sum_{i=1}^n \tilde{\beta}(i) \geq n, \text{ for some } n \geq s) \leq \mathrm{Prob}(\sum_{i=1}^n \tilde{\beta}(i) \geq s, \text{ for some } n \geq s)\\
&= \mathrm{Prob}(\sup_{n\geq s}\{\sum_{i=1}^n \tilde{\beta}(i)\} \geq s) \leq \mathrm{Prob}(\sup_{n\geq 0}\{\sum_{i=1}^n \tilde{\beta}(i)\} \geq s),
\end{aligned}$$

where we use the convention that the sum is 0 if the upper index in the sum is smaller than the lower index. This implies that

$$\tilde{T}^u \leq_{st} \sup_{n\geq 0}\{\sum_{i=1}^n \tilde{\beta}(i)\}. \tag{21}$$

Since the stochastic ordering is preserved under the composition of increasing functions,

$$\exp(\theta\tilde{T}^u) \leq_{st} \exp[\theta(\sup_{n\geq 0}\{\sum_{i=1}^n \tilde{\beta}(i)\})] \tag{22}$$

for $\theta > 0$. Replacing the supremum by the sum and taking expectations on both sides of (22), we obtain

$$Ee^{\theta\tilde{T}^u} \leq 1 + Ee^{\theta\tilde{\beta}(1)}\sum_{n=1}^\infty (\phi(\theta))^n \leq 1 + e^{\frac{\theta}{\epsilon_1\lambda}}\sum_{n=1}^\infty (\phi(\theta))^n \tag{23}$$

where $\phi(\theta) = E\exp(\theta\tilde{\beta}(2))$. Since $E\tilde{\beta}(2) = -1 < 0$ and $\phi(\theta) \leq \exp(\frac{\theta}{\epsilon_1\lambda}) < \infty$ for $\theta < \infty$, we always can find a $\theta > 0$ (small enough) such that $\phi(\theta) < 1$ (see Hajek [14], Lemma 2.2). Thus, the expectation in (23) is finite for some $\theta > 0$. This completes the argument for (P1).

(P2) Let $\tilde{T}^l = \sup\{n : n \le (\lambda - \epsilon)\tau(n)\}$. Now we derive the relation between T^l and $\tilde{T}^l$. If one can show that $\tau(\tilde{T}^l) \ge T^l$, it then follows from the definition of T^l and the monotonicity of $\bar{A}^r(t)$ in t that

$$\tilde{T}^l = \bar{A}^r(\tau(\tilde{T}^l)) \ge \bar{A}^r(T^l) \ge (\lambda - \epsilon)T^l - 1. \tag{24}$$

To show that $\tau(\tilde{T}^l) \ge T^l$, we condition on the event that $\tilde{T}^l = m$. Thus, for all $n \ge m$, $n \ge (\lambda - \epsilon)\tau(n)$. Suppose $\tau(n) \le t < \tau(n+1)$ for some $n \ge m$. Then we have that $\bar{A}^r(t) = n$ and that $n + 1 \ge (\lambda - \epsilon)\tau(n+1) \ge (\lambda - \epsilon)t$. This implies that $\bar{A}^r(t) \ge (\lambda - \epsilon)t - 1$. Thus, $\tau(\tilde{T}^l) \ge T^l$.

Now let

$$\epsilon_1 = \frac{\epsilon}{2\lambda(\lambda - \epsilon)} \text{ and } \tilde{\beta}(i) = \frac{1}{\epsilon_1}\Big(\beta(i) - \frac{1}{\lambda} - \epsilon_1\Big). \tag{25}$$

Using an argument similar to that in the proof for (P1), one can easily show that

$$\tilde{T}^l \le_{st} \sup_{n \ge 0}\{\sum_{i=1}^{n} \tilde{\beta}(i)\}. \tag{26}$$

and that

$$Ee^{\theta T^l} \le 1 + Ee^{\theta\tilde{\beta}(1)} \sum_{n=1}^{\infty} (\phi(\theta))^n \tag{27}$$

where $\phi(\theta) = E\exp(\theta\tilde{\beta}(2))$. Since $E\tilde{\beta}(2) = -1 < 0$ and $E\exp(\theta'\beta) < \infty$ for some $\theta' > 0$ by the assumption, there exists a $0 < \theta < \epsilon_1\theta'$ such that $\phi(\theta) < 1$ (Hajek [14], Lemma 2.2). To complete the argument, we also need to show that $E\exp(\theta\tilde{\beta}(1)) \le E\exp(\theta\beta(1)/\epsilon_1) < \infty$. Observe that the density of $\beta(1)$ is $\lambda\text{Prob}(\beta > s)$. Thus, applying integration by parts yields

$$Ee^{\frac{\theta}{\epsilon_1}\beta(1)} = \frac{\lambda\epsilon_1}{\theta} e^{\frac{\theta}{\epsilon_1}s}\text{Prob}(\beta > s)\Big|_0^\infty + \frac{\lambda\epsilon_1}{\theta} Ee^{\frac{\theta}{\epsilon_1}\beta}. \tag{28}$$

Since $E\exp(\theta'\beta) < \infty$, we have from Chernoff's bound (see for example Cover and Thomas [7], pp. 318) that $\lim_{s\to\infty}\exp(\theta s/\epsilon_1)\text{Prob}(\beta > s) \to 0$ for $\theta < \epsilon_1\theta'$. It then follows that

$$Ee^{\frac{\theta}{\epsilon_1}\beta(1)} = \frac{\lambda\epsilon_1}{\theta}(Ee^{\frac{\theta}{\epsilon_1}\beta} - 1) < \infty, \ \ \theta < \epsilon_1\theta'.$$

(P3) From (24), it is equivalent to showing that $E(\tilde{T}^l)^p < \infty$ under the assumption $E\beta^{p+1} < \infty$. From the proof for (P2), we have that

$$\tilde{T}^l \le_{st} \sup_{n\ge1}\{\sum_{i=1}^{n} \tilde{\beta}(i)\} \le \max(0, \tilde{\beta}(1)) + \sup_{n\ge1}\{\sum_{i=2}^{n} \tilde{\beta}(i)\}, \tag{29}$$

with $\tilde{\beta}(i)$ defined in (25). Note that $(x+y)^p \leq 2^p(x^p+y^p)$ (see Chow and Teicher, Lemma 4.2.3). Thus, in order to show $E(\tilde{T}^l)^p < \infty$, we need to show that both terms on the right hand side of (29) have finite p^{th} moments. Note that

$$E(\max(0,\tilde{\beta}(1)))^p \leq (\frac{1}{\epsilon_1})^p E(\beta(1))^p = (\frac{1}{\epsilon_1})^p \frac{\lambda E(\beta)^{p+1}}{p+1},$$

where the last equality is obtained by integration by parts (cf. (28)). Thus, $E(\max(0,\tilde{\beta}(1)))^p < \infty$ if $E(\beta)^{p+1} < \infty$. Since the second term in (29) is simply the maximum of a random walk with a negative drift, it then follows from Kiefer and Wolfowitz [17], Theorem 5 (see also Chow and Teicher [6], Corollary 10.4.3) that $E(\sup_{n\geq 1}\{\sum_{i=2}^{n}\tilde{\beta}(i)\})^p < \infty$ if $E(max(0,\tilde{\beta}(2)))^{p+1} < \infty$. Note that $\tilde{\beta}(2) \leq \frac{\beta(2)}{\epsilon_1}$. Thus, $E\beta^{p+1} < \infty$ implies $E(\sup_{n\geq 1}\{\sum_{i=2}^{n}\tilde{\beta}(i)\})^p < \infty$.

B Appendix B

In this appendix, we will prove Lemma 3.7 by a sample path argument. Let $\tau_1(n)$ be the n^{th} arrival epoch for both queues. Construct a stationary renewal process with interarrival times that have the same distribution as that of the service times and vacations in the queue with vacations. Let $\tau_2(n)$ be the n^{th} arrival epoch of this stationary renewal process. Let $\{\tau(n), n = 1,2,\ldots\}$ be the union of $\{\tau_1(n), n = 1,2,\ldots\}$ and $\{\tau_2(n), n = 1,2,\ldots\}$. Order $\{\tau(n), n = 1,2,\ldots\}$ in increasing sequence. The sequence $\{\tau(n), n = 1,2,\ldots\}$ is a sequence of event epochs. Let $q_n^a = q^a(\tau(n))$. Construct $q^a(t)$ inductively as follows:

(a1) $q^a(t) = q_{n-1}^a$, if $\tau(n-1) \leq t < \tau(n)$.

(a2) $q_n^a = q_{n-1}^a + 1$, if $\tau(n)$ is one of the points in $\{\tau_1(n), n = 1,2,\ldots\}$.

(a3) $q_n^a = \max(0, q_{n-1}^a - 1)$, if $\tau(n)$ is one of the points in $\{\tau_2(n), n = 1,2,\ldots\}$.

In words, the queue length $q^a(t)$ is not changed between two event epochs, increased by 1 if the event epoch is an arrival from arrival process, and decreased by 1 if the event epoch is an arrival from the service process and the queue is not empty.

Now we construct the queue length process $q^v(t)$ of the queue with vacations. Since we assume that the vacations have the same distribution as the service times in the queue with vacations, the union of all the departure epochs and the epochs that vacations end form a stationary renewal process. Thus, we can construct $q^v(t)$ analogous to the construction of $q^a(t)$ by introducing a flag $\zeta(t)$ that takes the value 1 if the server is on vacation and 0 otherwise. To simplify the notations, let $q_n^v = q^v(\tau(n))$ and $\zeta_n = \zeta(\tau(n))$.

(v1) $q^v(t) = q^v_{n-1}$ and $\zeta(t) = \zeta_{n-1}$, if $\tau(n-1) \le t < \tau(n)$.

(v2) $q^v_n = q^v_{n-1} + 1$ and $\zeta_n = \zeta_{n-1}$, if $\tau(n)$ is one of the points in $\{\tau_1(n), n = 1, 2, \ldots\}$.

(v3) $q^v_n = \max(q^v_{n-1} - (1 - \zeta_{n-1}), 0)$ if $\tau(n)$ is one of the points in $\{\tau_2(n), n = 1, 2, \ldots\}$

- **(v3a)** $\zeta_n = 0$ if $q^v_{n-1} > 0$ and $\zeta_{n-1} = 1$.
- **(v3b)** $\zeta_n = 1$ if $q^v_{n-1} = 1$ and $\zeta_{n-1} = 0$.
- **(v3a)** $\zeta_n = \zeta_{n-1}$ otherwise.

The only difference is (v3) where a departure occurs only when the queue is not empty and the server is not on vacation. The server comes back from vacation in (v3a) and the server takes a vacation in (v3b). The flag $\zeta(t)$ is not changed in all the other cases.

Now we claim that under our constructions $q^v(t) - (1 - \zeta(t)) = q^a(t)$ for all t. In other words, $q^v(t) = q^a(t) + 1$ if the server is not on vacation and $q^v(t) = q^a(t)$ if the server is on vacation. We will prove this by induction on $\tau(n)$ since there are no changes between two event epochs. It holds trivially at time 0 since both queues start from the same number of customers. If the queues start from empty, set $\zeta(0) = 1$. Otherwise, set $\zeta(0) = 0$. Assume that the induction hypothesis holds for $\tau(n-1)$. If $\tau(n)$ is an arrival from the arrival process, then $q^a_n = q^a_{n-1} + 1$, $q^v_n = q^v_{n-1} + 1$, and $\zeta_n = \zeta_{n-1}$. In conjunction with the induction hypothesis, it then follows that $q^v_n - (1 - \zeta_n) = q^a_n$. If $\tau(n)$ is an arrival from the service process, then we need to consider the following cases.

Case 1. $\zeta_{n-1} = 1$, $q^v_{n-1} = 0$:

In this case, the server takes another vacation. From the induction hypothesis, we have that $q^a_{n-1} = 0$. According to our construction, we have $q^a_n = q^v_n = 0$ and $\zeta_n = 1$. Thus, $q^v_n - (1 - \zeta_n) = q^a_n$.

Case 2. $\zeta_{n-1} = 1$, $q^v_{n-1} > 0$:

In this case, the server comes back from vacation and starts to work on the first customer in the queue. From the induction hypothesis, we have that $q^a_{n-1} = q^v_{n-1}$. According to our construction, we have $q^a_n = q^v_n - 1$ and $\zeta_n = 0$. Thus, $q^v_n - (1 - \zeta_n) = q^a_n$.

Case 3. $\zeta_{n-1} = 0$, $q^v_{n-1} = 1$:

In this case, the server completes the service of the last customer in the queue and takes a vacation. From the induction hypothesis, we have that $q^a_{n-1} = q^v_{n-1} - 1 = 0$. According to our construction, we have $q^a_n = q^v_n = 0$ and $\zeta_n = 1$. Thus, $q^v_n - (1 - \zeta_n) = q^a_n$.

Case 4. $\zeta_{n-1} = 0$, $q^v_{n-1} > 1$:

In this case, the server completes the service of a customer and continue to work on another customer in the queue. From the induction hypothesis, we have that $q^a_{n-1} = q^v_{n-1} - 1 > 0$. According to our construction, we still have $q^a_n = q^v_n - 1$ and $\zeta_n = 0$. Thus, $q^v_n - (1 - \zeta_n) = q^a_n$.

Since $0 \leq \zeta(t) \leq 1$, the statement in this lemma follows.

References

[1] S. Asmussen, *Applied Probability and Queues.* New York: J. Wiley & Sons, 1987.

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

[3] A.A. Borovkov, "Limit theorems for queueing networks. I", *Theory Probab. Appl.*, Vol. 31, pp. 413-427, 1986.

[4] C.S. Chang, "Stability, queue length and delay, Part I: Deterministic queueing networks," *IEEE Conference on Decision and Control*, Vol 1, pp. 999-1004, 1992.

[5] 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.

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

[7] T.M. Cover and J.A. Thomas, *Elements of Information Theory.* New York: Wiley & Sons, 1991.

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

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

[10] D.J. Daley and T. Rolski, "Finiteness of waiting-time moments in general stationary single-server queues," *Ann. Appl. Prob.*, 1992.

[11] D.J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes.* New York: Springer-Verlag, 1988.

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

[13] S.G. Foss. "Ergodicity of queueing networks," *Sibirskij Matematicheskij Ghurnal (Siberian Math. Journal)*, Vol. 32, No. 4, pp. 183-202, 1991 (in Russian). Translated in 1992.

[14] B. Hajek, "Hitting-time and occupation-time bounds implied by drift analysis with applications," *Adv. Appl. Prob.*, Vol. 14, pp. 502-525, 1982.

[15] T. Kamae, U. Krengel and G.L. O'Brien, "Stochastic inequalities on partially ordered spaces," *Ann. Prob.*, Vol. 5, pp. 899-912, 1977.

[16] J. Kiefer and J. Wolfowitz, "On the theory of queues with many servers," *Trans. Amer. Math. Soc.*, Vol. 78, pp. 1-18, 1955.

[17] J. Kiefer and J. Wolfowitz, "On the characteristics of the general queueing process with applications to random walk," *Ann. Math. Stat.*, Vol. 27, pp. 147-161, 1956.

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

[19] 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.

[20] J.F. Kurose, "On computing per-session performance bounds in high-speed multi-hop computer networks," to be presented in *SIGMETRICS and PERFORMANCE'92.*

[21] 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.

[22] S.P. Meyn and D. Down, "Stability of generalized Jackson networks," preprint, 1991.

[23] S.P. Meyn and R.L. Tweedie, "Criteria for stability of Markov processes I: Discrete time chain", *Adv. Appl. Prob.*, Vol. 24, pp. 542-574, 1992.

[24] S.P. Meyn and R.L. Tweedie, "Criteria for stability of Markov processes II: Continuous time processes and sample chains", preprint, 1990.

[25] Rolski, T.,"Queues with Nonstationary Arrivals," *Queueing Systems* 5, pp. 113-130, 1989.

[26] T. Rolski and R. Szekli, "Networks of work-conserving normal queues," *Applied Prob.-Computer Science. The Interface*, Vol. II. Birkhauser, 1982.

[27] T. Rolski and R. Szekli, "Stochastic ordering and thinning of point processes," *Stochastic Processes and Their Applications*, Vol. 37, pp. 299-312, 1991.

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

[29] J.G. Shanthikumar and D.D. Yao, "Stochastic monotonicity in general queueing networks," *J. Appl. Prob.*, Vol. 26, pp. 413-417, 1989.

[30] K. Sigman, "Notes on the stability of closed queueing networks," *J. Appl. Prob.*, Vol. 26, pp. 678-682, 1989. Also corrections in *J. Appl. Prob.*, Vol 27, pp. 735, 1990.

[31] K. Sigman, "The stability of open queueing networks," *Stochastic Processes and Their Applications*, Vol. 35, pp. 11-25, 1990.

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

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