# The Proof of a Folk Theorem on Queuing Delay with Applications to Routing in Networks

BRUCE HAJEK

*University of Illinois, Champaign-Urbana, Illinois*

Abstract. It is shown that among all arrival processes (not necessarily stationary or renewal type) for an exponential server queue with specified arrival and service rates, that the arrival process which minimizes the average delay and related quantities is the process with constant interarrival times. The proof is based on a convexity property of exponential server queues which is of independent interest. The folk theorem provides a lower bound, which is readily computable by existing methods, to the average delay in a network of queues under rather general routing disciplines. A sharper lower bound on average delay is provided for the special case of Generalized Round Robin routing for a Poisson arrival process.

Categories and Subject Descriptors: D.4.8 [**Operating Systems**]: Performance—*queuing theory; stochastic analysis*

General Terms: Performance, Theory

Additional Key Words and Phrases: Queues, routing, networks of queues, routing in queuing networks, packet communications

## 1. *Orientation and Statement of the Folk Theorem*

The purpose of this paper is to formulate and prove a certain folk theorem about queues. While the result is intuitively reasonable, its proof may not be so apparent. Roughly speaking, the theorem (stated as Theorem 1.1 below) says that among all arrival processes for an exponential server queue with specified average arrival and service rates, the arrival process which minimizes the average delay (and all moments of the delay, as well as other related quantities) is the process with constant interarrival times.

The proof of Theorem 1.1 is given in Section 3. The proof relies heavily on a convexity property of exponential server queues—namely, that the expected number of customers in an exponential server queue (as well as all higher moments) at a given time is a convex function of the set of previous interarrival times. This property is established in Section 2 and is of interest in its own right.

In Section 4 it is shown how Theorem 1.1 can be used to lower bound the average delay in a network of queues under a rather general class of routing strategies. This application, and indeed this whole paper, is motivated by the problem of minimum delay routing in a packet-switched communication network (see [1], [5–7], [9], [12], [16] and [20]).

This work was supported by JSEP contract N00014-79-C-0424.

Author's address: Coordinated Science Laboratory, University of Illinois, 1101 W. Springfield, Urbana, IL 61801.

The server considered in Theorem 1.1 is actually more general than an ordinary exponential server in that batches of customers can depart simultaneously. This generality is exploited in Section 5 to obtain a rather accurate bound on the delay in an exponential server queue when the arrival process is derived from a Poisson flow by a Generalized Round Robin routing policy.

The remainder of this section is divided into four subsections. In Section 1.1 the basic assumptions about the arrival and service processes are defined. In Section 1.2 it is shown that the folk theorem is easily established within the class of arrival processes with independent, identically distributed interarrival times, for then a variant of the well-known theory of G/M/1 queues applies. In Sections 1.3 and 1.4 the folk theorem is formulated and stated as Theorem 1.1 and its corollary.

1.1 Formulation of an Exponential Batch-Server Queue with a General Arrival Process. Consider a queue in which there is an exponential server with the modification that customers are served in batches. Specifically, when the queue is nonempty suppose that batches of service occur at a constant rate $\nu > 0$ and that the potential numbers of customers served in the batches are given by independent random variables with some common probability distribution $(q_k)_{k\geq 1}$. The number of customers $(N_t)_{t\geq 0}$ in such a queuing system subject to an arbitrary arrival random process can be described by the equation

$$\Delta N_t = \Delta A_t - \min(N_{t-}, \Delta D_t), \qquad t \geq 0; \qquad N_0 \quad \text{given},$$

where $\Delta N_t = N_t - N_{t-}$, and

(1) $(N_t)_{t\geq 0}$ is integer-valued, and has right-continuous-with-left-limits sample paths;
(2) $(A_t)_{t\geq 0}$ is a random counting process (i.e., $A_0 = 0$ and $(A_t)$ is integer-valued and has nondecreasing right-continuous sample paths);
(3) $(D_t)_{t\geq 0}$ is a compound Poisson counting process with jump rate $\nu > 0$ and jump size distribution $(q_k)$;
(4) $N_0$ is a nonnegative, integer-valued random variable.

Thus $A_t$ is the number of arrivals in the interval $(0, t]$, $D_t$ the number of *potential* departures in the interval $(0, t]$, and $N_0$ the initial number of customers in the system.

The average service rate is defined by

$$\mu = \nu \sum_{k=1}^{\infty} kq_k,$$

and may be infinite. The process $(D_t)$ is an independent-increment, pure jump Markov process with infinitesimal operator $\partial$ (i.e., backward Kolmogorov operator—see [3, Sections 15.4–15.5]) defined by

$$\partial f(l) = \nu \sum_{k=1}^{\infty} q_k(f(k+l) - f(l))$$

for bounded functions $f$ on $\mathbb{Z}_+$ (where $\mathbb{Z}_+$ is the set of nonnegative integers). For example,

$$\frac{d}{dt} Ef(D_t) = E\partial f(D_t) \tag{1.1}$$

for any bounded function $f$ on $\mathbb{Z}_+$, and if $p(k, t) = P(D_t = k)$, then

$$\frac{dp}{dt}(l, t) = \nu \sum_{k=1}^{\infty} q_k(p(l-k, t) - p(l, t)). \tag{1.2}$$

When $(q_k) = \delta_1$, where $\delta_1$ denotes the probability measure concentrated at $k = 1$, then $(D_t)$ is a Poisson process with rate $\mu = \nu$. In Section 5 an application is given in which a nondegenerate distribution $(q_k)$ naturally arises.

Throughout this section the following independence assumption is in effect:

$$(D_t)_{t\geq 0} \quad \text{is independent of} \quad (N_0, (A_t)_{t\geq 0}). \tag{1.3}$$

This assumption rules out "feedback" in the sense that it implies that given $N_0$ and $(A_s)_{0\leq s\leq t}$, the future of the arrival process $(A_s)_{s\geq t}$ is conditionally independent of $(N_s)_{0\leq s\leq t}$.

Define $\tau_k = \min\{t: A_t \geq k\}$. Thus $\tau_k$ is the time of arrival of the $k$th customer after time zero. Finally, define $\tilde{N}_k$ to be the number of customers in the system just before the $k$th arrival after time zero. If only one customer arrives (and no customer departs) at time $\tau_k$, then $\tilde{N}_k = N_{\tau_k} - 1$.

1.2 Special Case—Independent Identically Distributed Interarrival Times. We shall briefly consider the situation where the random variables $(\tau_{k+1} - \tau_k)_{k\geq 0}$ are independent and have a common distribution function $F$. The idea of this subsection is due to Ephremides [4]. Suppose that

$$E[\tau_{k+1} - \tau_k] = \lambda^{-1} \tag{1.4}$$

so that $\lambda$ is the average arrival rate, and define $\rho = \lambda/\mu$. By an easy extension of the well-known theory of G/M/1 queues [17], if $\rho < 1$, then

$$\lim_{k\to\infty} P(\tilde{N}_k = j) = (1 - x)x^j \tag{1.5}$$

where $x$ is the unique solution in the interval (0, 1) of the equation

$$F^*(-Q(x)) = x \tag{1.6}$$

where

$$Q(x) = \nu \sum_{k=1}^{\infty} (x^k - 1)q_k \tag{1.7}$$

and $F^*(s) = E \exp(-s(\tau_{k+1} - \tau_k))$ denotes the Laplace transform of $F$.

Among all distributions $F$ satisfying eq. (1.4), $x$ is minimized by $F = \delta_{\lambda^{-1}}$, the distribution concentrated at the single point $\lambda^{-1}$. That is, $x^* \leq x$ where $x^*$ is the unique number in the interval (0, 1) satisfying

$$\exp\left(\frac{Q(x^*)}{\lambda}\right) = x^*. \tag{1.8}$$

Indeed, by Jensen's inequality and the convexity of the function $\exp(-s\tau)$ in $\tau$ we have $\exp(-s/\lambda) \leq F^*(s)$ for $s \geq 0$. In particular, $\exp(Q(x)/\lambda) \leq F^*(-Q(x)) = x$. Since also $\exp(Q(1)/\lambda) = 1$ and the function $\exp(Q(y)/\lambda)$ is strictly convex, $\exp(Q(y)/\lambda) < y$ for $x < y < 1$ so that indeed $x^* \leq x$.

If $\Psi$ is any nondecreasing function on $\mathbf{Z}_+$, then the expectation of $\Psi$ with respect to the geometric distribution in eq. (1.5) is a nondecreasing function of $x$. Thus a consequence of eq. (1.5) and the fact that $x$ in eq. (1.5) satisfies $x \geq x^*$ is that if $\Psi$ is any nondecreasing function on $\mathbf{Z}_+$, then for any interarrival distribution $F$ with $\lambda$ fixed,

$$\lim_{k\to+\infty} E\Psi(\tilde{N}_k) \geq E^*\Psi \tag{1.9}$$

where

$$E^*\Psi = \begin{cases} \sum_{i=0}^{\infty} \Psi(i)(1 - x^*)(x^*)^i & \text{if} \quad \rho < 1, \\ \sup_i \Psi(i) & \text{if} \quad \rho \geq 1. \end{cases} \tag{1.10}$$

If $N_0$ is bounded, equality is achieved in (1.9) when $F = \delta_{\lambda^{-1}}$ or equivalently, when $\tau_k = k/\lambda$ for all $k$.

The main result of this paper (Theorem 1.1 below) is an extension of the inequality (1.9) to general arrival processes under the additional assumption that $\Psi$ is convex as well as nondecreasing.

1.3 General Arrival Processes. Since we will no longer assume that the arrival process $(A_t)$ is a renewal process, or that it is even asymptotically stationary, there is no longer a unique way to define the average arrival rate. We will thus consider the following two conditions, each of which roughly mean that the average arrival rate is at least $\lambda$:

$$\limsup_{n\to\infty} \frac{E\tau_n}{n} \leq \lambda^{-1} \tag{1.11}$$

$$\liminf_{T\to\infty} \frac{EA_T}{T} \geq \lambda. \tag{1.12}$$

Neither of these two conditions implies the other (see Appendix A). Note that $A_T/T$ is the time-averaged arrival rate over the fixed interval $(0, T]$, while $(\tau_n/n)^{-1}$ is the time-averaged arrival rate over the random interval $(0, \tau_n]$. Conditions (1.11) and (1.12) each reduce to the condition that $E(\tau_{k+1} - \tau_k) \leq \lambda^{-1}$ when the interarrival times are independent and identically distributed.

The main theorem will now be stated. It is proved in Section 3.

Theorem 1.1. *Let $\lambda > 0$ and define $\rho = \lambda/\mu$. Suppose that $\Psi$ is a nondecreasing convex function on $\mathbb{Z}_+$. Then*

(*a*) *Condition* (1.11) *implies that*

$$\liminf_{n\to\infty} E \frac{1}{n} \sum_{k=1}^{n} \Psi(\tilde{N}_k) \geq E^*\Psi;$$

(*b*) *Condition* (1.12) *implies that*

$$\liminf_{T\to\infty} \frac{E \sum_{k=1}^{A_T} \Psi(\tilde{N}_k)}{EA_T} \geq E^*\Psi;$$

*where $E^*\Psi$ is defined by eqs.* (1.10), (1.7), *and* (1.8).

1.4 Application to Waiting Times in Exponential Server Queues. In this subsection it is assumed that the queue has an exponential server of rate $\mu > 0$. Hence $(q_k) = \delta_1$, $Q(x) = \mu(x - 1)$, and eq. (1.8) for $x^*$ becomes

$$\exp \frac{x^* - 1}{\rho} = x^* \tag{1.13}$$

where $\rho = \lambda/\mu$. Now, let $W_k$ denote the waiting time in the queue of the $k$th customer to arrive, excluding the service time of the $k$th customer. The order of service can be arbitrary. Let $F_\rho$ denote the equilibrium waiting time distribution of a typical customer in a D/M/1 queue with first-come-first-serve (FCFS) service order. Explicitly [17],

$$F_\rho(y) = 1 - x^*\exp(-\mu(1 - x^*)y).$$

Corollary. *Suppose that $\Phi$ is a nondecreasing, convex function on $\mathbf{R}_+$. Then condition* (1.11) *implies that*

$$\liminf_{n\to\infty} E\frac{1}{n}\sum_{k=1}^{n} \Phi(W_k) \geq \int_0^\infty \Phi(x)F_\rho(dx),$$

*and condition* (1.12) *implies that*

$$\liminf_{T\to\infty} E\sum_{k=1}^{A_T} \frac{\Phi(W_k)}{EA_T} \geq \int_0^\infty \Phi(x)F_\rho(dx).$$

Proof. We will first prove the corollary under the assumption that the queue is initially empty and a FCFS service order is used. Define the function $\Psi$ on $\mathbf{Z}_+$ by $\Psi(n) = E[\Phi(W_k) \mid \tilde{N}_k = n]$. Then

$$E\Phi(W_k) = E\Psi(\tilde{N}_k) \tag{1.14}$$

and

$$\int_0^\infty \Phi(x)F_\rho(dx) = E^*\Psi. \tag{1.15}$$

It is shown in Appendix B that $\Psi$ is convex and nondecreasing on $\mathbf{Z}_+$. Therefore Theorem 1.1 applies and in view of eqs. (1.14) and (1.15) the conclusion of Theorem 1.1 coincides with that of the corollary. This establishes the corollary under the restriction that the queue is initially empty and a FCFS service order is used.

The corollary can then be deduced in the general case from the following fact. If the initial number of customers in the queue and all of the arrival and departure times are fixed, then for each $n$ the sum

$$\sum_{k=1}^{n} \Phi(W_k)$$

is minimized among all possible service orders by the FCFS order under the above restrictions on $\Phi$. (See [18]. Also, this fact is a consequence of an inequality due to Hardy, Littlewood, and Polya [8].) □

*Remark.* After the appearance of the original manuscript of this paper, P. Humblet [10] obtained a generalization of the corollary which includes systems with general independent identically distributed service times. (This also generalizes results in [15] and [2] in which the arrival processes are all assumed to be renewal type.) Humblet's proof is similar to our proof of Theorem 1.1 given in Section 3. Overall his proof is much simpler since for the problem he considers the convexity property (i.e., the analog of our Proposition 2.1) is immediate. Whether or not our Theorem 1.1 (which, in particular, allows batch service) can also be generalized to include general service distributions or, for example, multiple servers, remains open.

## 2. *A Convexity Property of Exponential Server Queues*

Fix a time $\tau$, and fix numbers $t_i$, $i \geq 1$, which give the lengths of the time intervals delineated by $\tau$ and the arrival instants before time $\tau$, counting backward. Thus, the arrival instants before time $\tau$ are $s_i = \tau - t_1 - \cdots - t_i$ for $i \geq 1$, and by convention we set $s_0 = \tau$. Also, fix a positive integer $m$ and suppose that the queue is empty just before time $s_m$. (This assumption is relaxed in the remarks following the proposition below.)

Assume that the server is governed by a compound Poisson potential departure process $(D_t)$ with jump rate $\nu$ and jump-size distribution $(q_k)$ as in Section 1. Hence, if $Y_i$ denotes the potential number of departures during the interval $[s_i, s_{i-1})$, then the distribution of the vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ depends on $\mathbf{t} = (t_1, t_2, \ldots, t_m)$ and is given by

$$Y_1, Y_2, \ldots, Y_m \quad \text{are independent;} \qquad P[Y_i = k] = p(k, t_i), \tag{2.1}$$

where $p(\cdot, t)$ is the probability distribution of $D_t$.

The number of customers in the queue just before time $s_i$ is the random variable $n_i(\mathbf{Y})$ where $n_i$, $0 \leq i \leq m$, are the functions of $\mathbf{y}$ in $\mathbb{Z}_+^m$ defined by

$$n_m(\mathbf{y}) = 0 \tag{2.2}$$

and

$$n_{i-1}(\mathbf{y}) = \max(n_i(\mathbf{y}) + 1 - y_i, 0) \qquad \text{for} \quad 1 \leq i \leq m. \tag{2.3}$$

In particular, $n_0(\mathbf{Y})$ is the number in the queue just prior to time $\tau$.

The main result of this section is the following proposition.

PROPOSITION 2.1. *For a function $\Psi$ on $\mathbb{Z}_+$ define the function $J_m$ on $\mathbb{R}_+^m$ by*

$$J_m(\mathbf{t}) = E\Psi(n_0(\mathbf{Y})). \tag{2.4}$$

*If $\Psi$ is nondecreasing then $J_m$ is nondecreasing in the component-wise ordering on $\mathbb{R}_+^m$. If $\Psi$ is nondecreasing and convex, then $J_m$ is convex on $\mathbb{R}_+^m$.*

*Remarks.* (1) If it is not assumed that the queue is empty just before time $s_m$, then $n_i(\mathbf{Y})$ (where $n_i(\mathbf{y})$ is defined by eqs. (2.2) and (2.3)) is less than or equal to the number in the queue just before time $s_i$.

(2) The proposition readily extends to cover the number in the queue at time $\tau$ as a function of the infinite collection $t_i$, $i \geq 1$ of interarrival times counted backward from time $\tau$. Begin with an alternative expression for $n_0(\mathbf{y})$ which can be derived from eqs. (2.2) and (2.3) by induction on $m$:

$$n_0(\mathbf{y}) = \max_{0 \leq i \leq m} i - \sum_{k=1}^{i} y_k,$$

where the sum is taken to be zero if $i = 0$. Thus,

$$J_m(t_1, \ldots, t_m) = E\Psi\left(\max_{0 \leq i \leq m} i - \sum_{k=1}^{i} Y_i\right) \tag{2.5}$$

where $(Y_i)$ satisfies (2.1). Now let $\Sigma$ be the set of infinite sequences $\mathbf{t} = (t_1, t_2, \ldots)$ where $0 \leq t_i \leq +\infty$. Suppose that $\Psi$ is nondecreasing. Then expression (2.5) shows that as a function on $\Sigma$, $J_m$ is nondecreasing in $m$ for each $\mathbf{t}$. Hence we may define

$$J(\mathbf{t}) = \lim_{m \to \infty} J_m(t_1, \ldots, t_m) \tag{2.6}$$

with the provision that $J(\mathbf{t})$ may equal $+\infty$. Letting $m$ tend to infinity in eq. (2.5) and using Lebesgue's Monotone Convergence Theorem implies that

$$J(\mathbf{t}) = E\Psi\left(\sup_{i \geq 0} i - \sum_{k=1}^{i} Y_k\right) \tag{2.7}$$

with the provision that $\Psi(+\infty) = \sup\{\Psi(k):k \geq 0\}$. Equation (2.6) provides a direct probablistic interpretation of $J$. The function $J$ is the analog corresponding to $m = +\infty$ of the function $J_m$. Proposition 2.1 extends to $J$—that is, $J$ is nonincreasing in the componentwise order on $\Sigma$, and if $\Psi$ is convex, then $J$ is convex on $\Sigma$. Indeed, these properties are true since $J$ is the limit of functions with such properties.

In the special case that $t_i = \lambda^{-1}$ for all $i$, the random variable

$$\sup_{i\geq 0} i - \sum_{k=1}^{i} Y_k \tag{2.8}$$

has the distribution of the number in a D/M/1 queue just before a typical arrival time. Therefore eqs. (2.6) and (2.7) imply that

$$\lim_{m\to\infty} J_m(\lambda^{-1}, \lambda^{-1}, \ldots, \lambda^{-1}) = E^*\Psi. \tag{2.9}$$

Proposition 2.1 will be proved after we present some preliminary notation and lemmas. Let $\alpha_i$ denote the map from $\mathbb{Z}^m$ to $\mathbb{Z}^m$ which increases the $i$th coordinate by one. Thus

$$\alpha_i(y_1, \ldots, y_m) = (y_1, \ldots, y_{i-1}, y_i + 1, y_{i+1}, \ldots, y_m).$$

Then for $k \geq 0$ and $1 \leq i \leq m$ define the operator $\Delta_{i,k}$ acting on functions $f$ on $\mathbb{Z}_+^m$ by $\Delta_{i,k}f(\mathrm{y}) = f(\alpha_i^k \mathrm{y}) - f(\mathrm{y})$. Finally, for $\mathrm{y} \in \mathbb{Z}_+^m$ and $1 \leq i \leq m$ let

$$g_i(\mathrm{y}) = \min\{n_a(\mathrm{y}):0 \leq a < i\}. \tag{2.10}$$

Note that $g_1(\mathrm{y}) = n_0(\mathrm{y})$. Moreover, $g_i(\mathrm{y})$ is the largest amount by which $n_0(\mathrm{y})$ can be decreased by increasing $y_i$ (and leaving the other coordinates of y fixed).

LEMMA 2.2. *For* $1 \leq i \leq j \leq m$,

$$n_0(\alpha_i^k \mathrm{y}) = n_0(\mathrm{y}) - \mathit{min}(k, g_i(\mathrm{y})), \tag{2.11}$$

$$g_i(\alpha_j^l \mathrm{y}) = g_i(\mathrm{y}) - \mathit{min}(l, g_j(\mathrm{y})), \tag{2.12}$$

*and*

$$n_0(\alpha_i^k \alpha_j^l \mathrm{y}) = n_0(\mathrm{y}) - \mathit{min}(k + l, k + g_j(\mathrm{y}), g_i(\mathrm{y})). \tag{2.13}$$

PROOF. Equation (2.11) is an easy consequence of the interpretation of $g_i$. To prove eq. (2.12), consider what happens as the number of potential services $y_j$ is increased one at a time. Call an additional potential service effective if it causes $n_0$ to decrease by one. Now, each potential service corresponding to increasing $y_j$ by one which is effective will decrease $n_b$ by one for all $b$ with $0 \leq b < j$, and hence $g_i$ will decrease by one. On the other hand, if such an additional potential service is not effective, then (prior to the addition of the service) either $g_i = 0$ or $n_b = 0$ for some $b$ with $i \leq b < j$. Consequently $g_i$ is unchanged by an increase in $\mathrm{y}_j$ which is not effective. This establishes eq. (2.12).

Application of eq. (2.11) with y replaced by $\alpha_j^l \mathrm{y}$ yields that

$$n_0(\alpha_i^k \alpha_j^l \mathrm{y}) = n_0(\alpha_j^l \mathrm{y}) - \min(k, g_i(\alpha_j^l \mathrm{y})). \tag{2.14}$$

Finally, using eq. (2.11) with $i$ replaced by $j$ and eq. (2.12) in the right side of (2.14) yields eq. (2.13). □

LEMMA 2.3. *For fixed* y $\in \mathbb{Z}_+^m$, *let* $n_0$ *and* $g_i$ *denote* $n_0(\mathrm{y})$ *and* $g_i(\mathrm{y})$, *respectively. Then for* $1 \leq i \leq j \leq m$,

$$\Delta_{i,k}\Psi(n_0(\mathrm{y})) = \Psi(n_0 - \mathit{min}(k, g_i)) - \Psi(n_0) \tag{2.15}$$

*and*

$$\Delta_{i,k}\Delta_{j,l}\Psi(n_0(\mathrm{y})) = \Psi(n_0) - \Psi(n_0 - min(k, g_i)) \\ - \Psi(n_0 - min(l, g_j)) + \Psi(n_0 - min(k + l, k + g_j, g_i)). \quad (2.16)$$

PROOF. Lemma 2.3 is an immediate consequence of eqs. (2.11) and (2.13) and the definitions of $\Delta_{i,k}$ and $\Delta_{j,l}$. □

The identity given in the next lemma provides an alternative expression for the right side of eq. (2.16).

LEMMA 2.4. *Adopt the convention that $\Psi(k) = \Psi(0)$ for all $k \leq 0$. Then for $0 \leq s \leq r \leq n_0$ and $k, l \geq 0$,*

$$\Psi(n_0) - \Psi(n_0 - min(k, r)) - \Psi(n_0 - min(l, s)) \\ + \Psi(n_0 - min(k + l, k + s, r)) \quad (2.17)$$

*is equal to (using $\{a < r, s \leq b\}$ to denote $\{a < r \leq b$ and $a < s \leq b\}$)*

$$\sum_{a=0}^{l-1} I\{a < r, s\}[\Psi(n_0 - a) - \Psi(n_0 - a - 1) - \Psi(n_0 - k - a) \\ + \Psi_0(n - k - a - 1)] \\ + \sum_{a=0}^{l-1} I\{a < r, s \leq a + k\}[\Psi(n_0 - k - a) - \Psi(n_0 - k - a - 1)]. \quad (2.18)$$

PROOF. The first sum in expression (2.18) is equal to

$$\sum_{a=0}^{l-1} I\{a < s\}[\Psi(n_0 - a) - \Psi(n_0 - a - 1)] \\ - \sum_{a=0}^{l-1} I\{a < s\}[\Psi(n_0 - k - a) - \Psi(n_0 - k - a - 1)]$$

which, since the sums telescope, is equal to

$$\Psi(n_0) - \Psi(n_0 - \min(l, s)) - [\Psi(n_0 - k) - \Psi(n_0 - k - \min(l, s))]. \quad (2.19)$$

The second sum in expression (2.18) is equal to

$$\sum_{a=0}^{l-1} I\{\min(s, (r - k)^+) \leq a < s\}[\Psi(n_0 - k - a) - \Psi(n_0 - k - a - 1)] \\ = -\Psi(n_0 - k - \min(l, s)) + \Psi(n_0 - k - \min(l, s, (r - k)^+)). \quad (2.20)$$

Combining expressions (2.19) and (2.20) yields that (2.18) is equal to

$$\Psi(n_0) - \Psi(n_0 - \min(l, s)) - \Psi(n_0 - k) + \Psi(n_0 - k - \min(l, s, (r - k)^+)). \quad (2.21)$$

Now if $k \geq r$, the sum of the last two terms in expression (2.21) is zero, so the expression is unchanged if $k$ is replaced by $\min(k, r)$. Thus, (2.18) is equal to

$$\Psi(n_0) - \Psi(n_0 - \min(l, s)) - \Psi(n_0 - \min(k, r)) \\ + \Psi(n_0 - \min(k, r) - \min(l, s, (r - k)^+)).$$

By considering separately the cases $k \leq r$ and $k > r$, this expression is readily seen to be equal to expression (2.17). □

LEMMA 2.5. *Suppose $\Psi$ is convex and nondecreasing on $\mathbb{Z}^+$. Then for all $\mathrm{y} \in \mathbb{Z}_+^m$ and integers $k, l \geq 0$, the matrix*

$$\{(\Delta_{i,k}\Delta_{j,l} + \Delta_{i,l}\Delta_{j,k})\Psi(n_0(\mathrm{y}))\}_{1 \leq i, j \leq m} \quad (2.22)$$

*is positive semidefinite.*

**Proof.** Let $n_0$ and $g_i$ denote $n_0(\mathbf{y})$ and $g_i(\mathbf{y})$, respectively. If $i \le j$, then $0 \le g_j \le g_i \le n_0$, so that Lemma 2.4 with $r = g_i$ and $s = g_j$ can be applied to the right side of eq. (2.16). Thus, if $i \le j$, then

$$\begin{aligned}
&(\Delta_{i,k}\Delta_{j,l} + \Delta_{i,l}\Delta_{j,k})\Psi(n_0(\mathbf{y}))\\
&\quad = \sum_{a=0}^{l-1} I\{a < g_i, g_j\}[\Psi(n_0 - a) - \Psi(n_0 - a - 1) - \Psi(n_0 - k - a)\\
&\qquad\qquad + \Psi(n_0 - k - a - 1)]\\
&\quad + \sum_{a=0}^{k-1} I\{a < g_i, g_j\}[\Psi(n_0 - a) - \Psi(n_0 - a - 1) - \Psi(n_0 - l - a)\\
&\qquad\qquad + \Psi(n_0 - l - a - 1)]\\
&\quad + \sum_{a=0}^{l-1} I\{a < g_i, g_j \le a + k\}[\Psi(n_0 - k - a) - \Psi(n_0 - k - a - 1)]\\
&\quad + \sum_{a=0}^{k-1} I\{a < g_i, g_j \le a + l\}[\Psi(n_0 - l - a) - \Psi(n_0 - l - a - 1)].
\end{aligned} \tag{2.23}$$

Since each side of eq. (2.23) is symmetric in $i$ and $j$, the equation is valid for all $i, j$ with $1 \le i, j \le m$.

Now each matrix of the form $(I\{a < g_i, g_j \le b\})_{1 \le i, j \le m}$ is positive semidefinite since it can be written as $vv^{\mathrm{T}}$ where $v$ is the $m$-vector with $v_i = I\{a < g_i \le b\}$. Thus eq. (2.23) and the properties of $\Psi$ show that the matrix (2.22) is a linear combination with nonnegative coefficients of symmetric positive semidefinite matrices. Since a symmetric $m \times m$ matrix $A$ is positive semidefinite if and only if $v^{\mathrm{T}}Av$ is nonnegative for all $m$-vectors $v$ it is clear that the set of such matrices is closed under addition. This implies the desired result. □

**Proof of Proposition 2.1.** Define the operator $\partial_i$ acting on bounded functions $f$ on $\mathbf{Z}_+^m$ by

$$\partial_i f(y) = \nu \sum_{k=1}^{\infty} q_k \Delta_{i,k} f(y). \tag{2.24}$$

Then $\partial_i$ is the same as $\partial$ applied to $f(\mathbf{y})$ as a function of $y_i$ for the other coordinates of $\mathbf{y}$ fixed. Since the random variables $Y_1, \ldots, Y_m$ are independent, repeated use of eq. (1.1) yields that

$$\frac{\partial J_m}{\partial t_i}(\mathbf{t}) = E\partial_i\Psi(n_0(\mathbf{Y}))$$

and

$$\frac{\partial^2 J_m}{\partial t_i \partial t_j}(\mathbf{t}) = E\partial_i\partial_j\Psi(n_0(\mathbf{Y})).$$

Thus, by eq. (2.24),

$$\frac{\partial J_m}{\partial t_i}(\mathbf{t}) = E\nu \sum_{k=1}^{\infty} q_k \Delta_{i,k}\Psi(n_0(\mathbf{Y})) \tag{2.25}$$

and

$$\frac{\partial^2 J_m}{\partial t_i \partial t_j}(\mathbf{t}) = E\nu^2 \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} q_k q_l \Delta_{i,k}\Delta_{j,l}\Psi(n_0(\mathbf{Y})). \tag{2.26}$$

By the symmetry in $k$ and $l$ of the sum in eq. (2.26),

$$\frac{\partial^2 J_m}{\partial t_i \partial t_j}(\mathbf{t}) = E\frac{\nu^2}{2}\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} q_k q_l[\Delta_{i,k}\Delta_{j,l} + \Delta_{i,l}\Delta_{j,k}]\Psi(n_0(\mathbf{Y})). \tag{2.27}$$

By eq. (2.15) each term in the sum on the right side of eq. (2.25) is nonpositive if $\Psi$ is nondecreasing. In this case, therefore, $J_m$ is nonincreasing in $\mathbf{t}$.

Now suppose that $\Psi$ is convex and nondecreasing. By Lemma 2.5 and the representation (2.27) the Hessian matrix $\nabla^2 J_m$ of second partial derivatives is a limit of linear combinations with nonnegative coefficients of symmetric positive semidefinite matrices. Since the set of symmetric positive semidefinite matrices is closed under addition and limits (see proof of Lemma 2.5), $\nabla^2 J_m$ is thus also positive semidefinite. Hence, $J_m$ is convex (see [14, p. 27, Th. 4.5]). □

## 3. *The Proof of Theorem* 1.1

First the basic idea of the proof of Theorem 1.1 will be explained and then the detailed proof will follow. At the time of each arrival the expected value of the function $\Psi$ of the number of customers in the system is a function (given by $J$ in Section 2) of the interarrival times for the previous arrivals. By choosing a customer at random a distribution is induced on the vector of past interarrival times. The fact that the arrival rate is at least $\lambda$ places a constraint on this distribution. Specifically, it implies that the average interarrival time is at most $\lambda^{-1}$. Thus Jensen's inequality and the convexity property of $J$ established in Proposition 2.1 yield the desired result. We shall now begin the actual proof.

Assume the notation of Section 1 and suppose that $\Psi$ is a nondecreasing convex function on $\mathbb{Z}_+$. For fixed $k$, the random variables $\tau_k - \tau_{k-1}, \tau_{k-1} - \tau_{k-2}, \ldots$ are the interarrival times, counted backward from time $\tau_k$. Since $\tilde{N}_k$ is the number in the queue just before time $\tau_k$, Proposition 2.1 (and the first remark following it) yield that for all integers $m$ and $k$ with $k \geq m + 1$,

$$E(\Psi(\tilde{N}_k)\,|\,A_t, t \geq 0, N_0) \geq J_m(\tau_k - \tau_{k-1}, \tau_{k-1} - \tau_{k-2}, \ldots, \tau_{k-m+1} - \tau_{k-m}) \tag{3.1}$$

where $J_m$ is defined as in Section 2.

We now turn specifically to the proof of part (a) of Theorem 1.1. Thus, assume that condition (1.11) is true. Now taking expectations of each side of eq. (3.1) yields that

$$E\Psi(\tilde{N}_k) \geq EJ_m(\tau_k - \tau_{k-1}, \tau_{k-1} - \tau_{k-2}, \ldots, \tau_{k-m+1} - \tau_{k-m}).$$

Hence

$$E\sum_{k=m+1}^{n} \Psi(\tilde{N}_k)$$

$$\geq (n-m)E\frac{1}{n-m}\sum_{k=m+1}^{n} J_m(\tau_k - \tau_{k-1}, \tau_{k-1} - \tau_{k-2}, \ldots, \tau_{k-m+1} - \tau_{k-m}). \tag{3.2}$$

Now for fixed $k$, the $\mathbb{R}_+^m$-valued vector $(\tau_k - \tau_{k-1}, \ldots, \tau_{k-m+1} - \tau_{k-m})$ is random and is thus by definition a function on the underlying probability space $\Omega$. By letting $k$ vary, the vector can be viewed as an $\mathbb{R}_+^m$-valued function on the set $\Omega \times \{k : m + 1 \leq k \leq n\}$, and the symbols

$$E\frac{1}{n-m}\sum_{k=m+1}^{n}$$

represent an expectation with respect to a probability measure on this set. Thus, since $J_m$ is convex on $\mathbb{R}_+^m$, Jensen's inequality [3, p. 80] can be used to bound the right side of inequality (3.2) from below to yield that

$$\begin{aligned} &E \sum_{k=m+1}^{n} \Psi(\tilde{N}_k) \\ &\quad \geq (n-m)J_m\left(E\frac{1}{n-m}\sum_{k=m+1}^{n}(\tau_k-\tau_{k-1}, \tau_{k-1}-\tau_{k-2}, \ldots, \tau_{k-m+1}-\tau_{k-m})\right) \\ &\quad = (n-m)J_m\left(E\frac{1}{n-m}(\tau_n-\tau_m, \tau_{n-1}-\tau_{m-1}, \ldots, \tau_{n-m+1}-\tau_1)\right) \\ &\quad \geq (n-m)J_m\left(E\frac{1}{n-m}(\tau_n, \tau_n, \ldots, \tau_n)\right). \end{aligned}$$

The last inequality is a consequence of the fact that $J_m$ is a nonincreasing function on $\mathbb{R}_+^m$ (see Proposition 2.1) and $\tau_n \geq \tau_{n-j} - \tau_{m-j}$ for $0 \leq j \leq m-1$. Now divide through by $n$ and let $n$ tend to infinity (for $m$ fixed) to yield that (using condition (1.11))

$$\liminf_{n\to\infty} E\frac{1}{n}\sum_{k=1}^{n}\Psi(\tilde{N}_k) \geq J_m(\lambda^{-1}, \lambda^{-1}, \ldots, \lambda^{-1}).$$

Then by (2.9) as $m$ tends to infinity the right side of this inequality converges to $E^*\Psi$. This establishes part (*a*) of Theorem 1.1.

We shall now turn to the proof of part (b) of Theorem 1.1, so assume that condition (1.12) is true. Now multiplying each side of eq. (3.1) by $I\{\tau_k \leq T\}$ and taking expectations yields that

$$E\Psi(\tilde{N}_k)I\{\tau_k \leq T\} \geq E(J_m(\tau_k-\tau_{k-1}, \ldots, \tau_{k-m+1}-\tau_{k-m})I\{\tau_k \leq T\}).$$

Forming the sum of each side of this inequality over $k$ from $m+1$ to infinity yields that

$$E\sum_{k=m+1}^{A_T}\Psi(\tilde{N}_k) \geq E\sum_{k=m+1}^{\infty} J_m(\tau_k-\tau_{k-1}, \ldots, \tau_{k-m+1}-\tau_{k-m})I\{\tau_k \leq T\}. \tag{3.3}$$

Now

$$E\sum_{k=m+1}^{\infty} I\{\tau_k \leq T\} = E\max(A_T-m, 0),$$

so that using the convexity of $J_m$ and Jensen's inequality to lower-bound the right side of inequality (3.3) yields that

$$\begin{aligned} &E\sum_{k=m+1}^{A_T}\Psi(\tilde{N}_k) \\ &\quad \geq E\max(A_T-m,0)J_m\left(\frac{E\sum_{k=m+1}^{\infty}(\tau_k-\tau_{k-1}, \ldots, \tau_{k-m+1}-\tau_{k-m})I\{I_k \leq T\}}{E\max(A_T-m,0)}\right) \\ &\quad \geq (EA_T-m)J_m\left(\frac{(T, T, \ldots, T)}{EA_T-m}\right) \end{aligned} \tag{3.4}$$

where we have assumed that $T$ is so large that $EA_T - m \geq 0$. The final inequality is a consequence of the fact that $J_m$ is a nonincreasing function on $\mathbb{R}_+^m$ and the inequalities

$$E\max(A_T-m, 0) \geq EA_T - m$$

and

$$\sum_{k=m+1}^{\infty} (\tau_{k-j} - \tau_{k-j-1}) I\{\tau_k \le T\} \le T \qquad \text{for} \quad 0 \le j \le m-1.$$

Dividing each side of inequality (3.4) by $EA_T$ and letting $T$ tend to infinity (for $m$ fixed) yields that (using condition (1.12))

$$\liminf_{T\to\infty} E \sum_{k=1}^{A_T} \frac{\Psi(\tilde{N}_k)}{EA_T} \ge J_m(\lambda^{-1}, \lambda^{-1}, \ldots, \lambda^{-1}).$$

Finally, by (2.9) as $m$ tends to infinity the right side of this inequality converges to $E^*\Psi$, and the proof of Theorem 1.1 is complete. □

## 4. *Application to Routing in Queuing Networks*

The purpose of this section is to apply Theorem 1.1 to provide a lower bound to the mean message delay in a network of exponential server queues under nonfeedback, loop-free routing strategies. First the network model will be described. Consider a network of $M$ queues such that the service time distribution of the $k$th queue is exponential with mean $\mu_k^{-1}$. Let customers enter the $k$th queue from outside the network at an average rate $\gamma_k$, and suppose that there is an $M \times M$ matrix $R = \{r_{l,k}\}$ (called the routing matrix) such that, over a long period of time, a fraction $r_{l,k}$ of the customers which depart from queue $l$ are next sent to queue $k$. Typically $r_{l,k} \ne 0$ only if $(l, k) \in \mathscr{C}$ for some subset $\mathscr{C} \subset \{1, \ldots, M\} \times \{1, \ldots, M\}$ denoting connectivity.

We shall assume that the network is stable in the sense that the average rate of flow out of each queue is equal to the average rate of flow into the queue. Then the total average rates of flow $\lambda_1, \ldots, \lambda_M$ into the $M$ queues are determined by the conservation of flow equations

$$\lambda_k = \gamma_k + \sum_l \lambda_l r_{l,k}.$$

A problem often faced in the design and operation of such a queuing network is to choose appropriate parameter values (e.g., the routing matrix) in order to minimize the average delay per customer in passing through the network. Hence it is important to calculate or at least to bound the average delay. One solution to this problem is to assume that the incoming streams of customers are given by independent Poisson processes, and to assume that an "independent splitting" routing strategy, described next, is used. (Another routing strategy is described in the next section.) Under an independent splitting routing strategy, a customer exiting from queue $l$ decides to join queue $k$ next with probability $r_{l,k}$, and the decision of which queue to join next is made independently of the past history of the entire network. Then by a well-known theorem of Jackson [9] the equilibrium distribution can be explicitly described (it has the "product form"), and the average delay per customer (service times included) is given by

$$D(\lambda) = \frac{1}{\gamma} \sum_k d\left(\frac{\lambda_k}{\mu_k}\right) \tag{4.1}$$

where $\gamma = \sum_k \gamma_k$ and $d(\rho) = \rho/(1-\rho)$.

Various algorithms have been proposed for choosing the routing matrix $R$ to minimize $D$ in (4.1) for a given set of input rates $(\gamma_k)$. (See, e.g., [1, 7, 9, 12, 16].) It is often the case, however, that some routing strategies other than independent splitting achieve a smaller average delay per customer. For example, for routing a single stream of traffic through two parallel exponential server queues, it is shown in [5] that the routing strategy which minimizes the average delay per customer is the Round Robin strategy which sends every other arrival to one queue. For more general networks it appears very difficult to find the minimum delay routing strategy, although it is clear that in most cases the optimal routing strategy is not an independent splitting type [5, 20].

Hence it is important to have a lower bound on the average delay achievable by general routing strategies. Such a bound is implied by Theorem 1.1. The bound is valid even if the incoming traffic streams are not Poisson processes, and the routing strategies need not be time-invariant.

In order to apply Theorem 1.1, assume that the total average arrival rate (in the sense of condition (1.11) or condition (1.12)) of customers into the $k$th queue is at least $\lambda_k$. Further, we must require that the independence assumption (1.3) holds for each of the queues in the network. This assumption rules out feedback routing strategies in which the number of customers in the queues downstream is fed back to the routing mechanism at a given queue (such as in [6]). Moreover, in most instances, the assumption is likely to be invalid if it is possible for customers to travel in loops—thereby visiting a given queue more than once.

Under these assumptions, the corollary to Theorem 1.1 yields that the average waiting time (including service time) in the $k$th queue for customers passing through the $k$th queue is at least $1/(1 - x^*)\mu_k$ where $x^*$ satisfies eq. (1.13) with $\rho = \lambda_k/\mu_k$. (Once again, average waiting time here is in the sense of Theorem 1.1.) Thus by Little's result, the average number of customers in the $k$th queue is at least $\lambda_k/(1 - x^*)\mu_k = \tilde{d}(\rho)$ where (using eq. (1.13) for $x^*$), $\tilde{d}$ is the function implicitly defined by

$$\rho = \tilde{d}(\rho)\left(1 - \exp\left[\frac{-1}{\tilde{d}(\rho)}\right]\right), \qquad 0 \leq \rho < 1. \tag{4.2}$$

Hence, again by Little's result (applied to the network as a whole), the average waiting time per packet in the network is lower bounded by

$$\frac{1}{\gamma}\sum_k \tilde{d}\left(\frac{\lambda_k}{\mu_k}\right) \tag{4.3}$$

where $\tilde{d}$ is defined by eq. (4.2). Now $\tilde{d}$ is convex and increasing on (0, 1) and $\tilde{d}(\rho)$ tends to infinity as $\rho$ approaches one. Since the expression (4.3) has the same form as $D$ in eq. (4.1), existing algorithms [1, 9] can be used to minimize the expression (4.3) over the parameters of interest, yielding an absolute lower bound to average delay for all loop-free, nonfeedback routing strategies with given external arrival rates $(\gamma_k)$.

## 5. *Application to Generalized Round Robin Splitting of a Poisson Flow*

Consider a flow of customers who arrive according to a Poisson process $(\bar{A}_t)_{t\geq 0}$ with rate $\alpha \geq 0$ and which are routed into one of $M$ parallel queues. A Round Robin (RR) routing policy is characterized by the fact that any $M$ consecutive arrivals are routed to the $M$ distinct queues. In order to split the traffic asymmetrically, a Generalized Round Robin (GRR) policy could be used. Under a GRR routing policy, customers are routed solely on the basis of their order of arrival. More precisely, a GRR routing policy is specified by an infinite sequence $(s_k)_{k\geq 0}$ (called a routing or splitting sequence) such that $1 \leq s_k \leq M$ and $s_k$ indicates the queue to which the $k$th arrival will be routed.

Suppose that a GRR routing policy is used and consider one of the $M$ queues, say queue number one. Suppose the queue has an exponential server of rate $\mu$, and let $(\bar{D}_t)$ be a Poisson process of rate $\mu$, independent of $(\bar{A}_t)$, which models potential departures at queue one. Let $u_i$ denote the index, among arrivals routed to all queues, of the $i$th customer to be routed to queue one, and let $A_k$ denote the number among the first $k$ arrivals which are routed to queue one. Note that $(u_i)$ and $(A_k)$ are deterministic sequences which are simply determined by the routing sequence. For example, the sequences $u_n = nM$ and $A_k = [k/M]$ (where $[x]$ denotes the largest integer less than or equal to $x$) arise from a (nongeneralized) RR strategy.

Let $\tilde{N}_i$ denote the number of customers in queue one just prior to the $i$th arrival at queue one. The number $p$ represents the fraction of arrivals which are routed to queue number one. The purpose of this section is to prove the following proposition.

PROPOSITION 5.1. *Let $\Psi$ be a nondecreasing, convex function on $\mathbb{Z}_+$. Then*

(*a*) *If* $\limsup_{n\to\infty} \frac{u_n}{n} \le p^{-1}$, *then* $\liminf_{n\to\infty} E\frac{1}{n}\sum_{i=1}^{n} \Psi(\tilde{N}_i) \ge E^*\Psi$;

(*b*) *If* $\liminf_{k\to\infty} \frac{A_k}{k} \ge p$, *then* $\liminf_{k\to\infty} E\sum_{i=1}^{A_k} \frac{\Psi(\tilde{N}_i)}{EA_k} \ge E^*\Psi$;

*where $E^*\Psi$ is defined by eq.* (1.10) *with $\rho = \alpha p/\mu$ and if $\rho < 1$, then $x^*$ is the unique solution in the interval* (0, 1) *of the equation*

$$\left(\frac{\alpha}{(1 - x^*)\mu + \alpha}\right)^{1/p} = x^*. \tag{5.1}$$

*The inequalities in* (*a*) *and* (*b*) *are all equalities if $u_n = nK$ and $p = 1/K$ for some positive integer $K$.*

*Remark.* Many GRR routing disciplines are attractive because they are as easy to implement as independent splitting (which was described in Section 4) and yet often provide significantly smaller queuing delays [20, 5]. For example, we conjecture that for a given probability $p$ a routing sequence such that $A_k = [pk]$ yields the smallest mean delay at queue one among all deterministic routing sequences for which either condition (a) or (b) of Proposition 5.1 are satisfied.[1] (Curiously enough, although such conditions are compatible for two parallel queues they can be incompatible for three or more queues—suppose for example that the stream is to be split three ways in proportions (1/2, 1/3, 1/6).) Proposition 5.1 establishes this conjecture in case $p = 1/K$ for some integer $K$. If $p$ is rational this rule can be realized by periodic routing sequences. If the routing sequence is periodic, then the limiting distribution of $\tilde{N}_i$ can be conveniently computed by the method of Markov processes with phases (see Neuts [13]). This was carried out in [19] and some results are pictured in Figure 1. The bound given by Proposition 5.1 is nearly reached for some periodic GRR sequences.

PROOF. Extend the sequence $(A_k)$ to a (deterministic) continuous-parameter process $(A_u)_{u\in\mathbb{R}_+}$ by defining $A_u = A_{[u]}$. Note that the sequence $(u_k)$ consists of the jump times of $(A_u)$. The idea of the proof is to show that $\tilde{N}_i$ can be viewed as the number of customers in a queue just before the $i$th arrival when the arrival process is $(A_u)$ and the server is governed by a certain compound Poisson process, representing potential departures. In essence, the trick is to view the potential service process $(\bar{D}_t)$ at queue one and the total arrival process $(\bar{A}_t)$ on a new time scale. One unit of time

[1] This conjecture has been established since this paper was originally written; information regarding its publication is available from the author

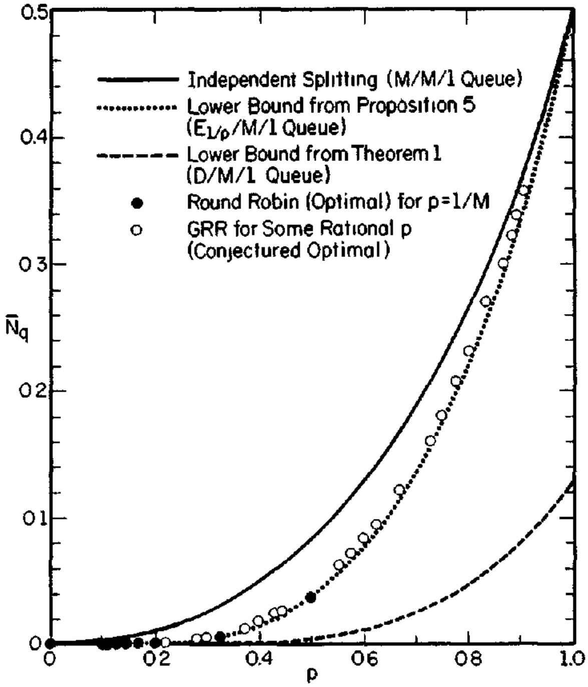

FIG. 1. The mean number in an exponential server queue of rate $\mu$ (excluding possible customer in service) versus $p$, where $p$ is the fraction of a Poisson stream of rate $\lambda$ which is routed into the queue The ratio $\rho = \lambda/\mu$ is fixed at 0.5.

on the new scale (which we parameterize by $u$) corresponds to one interarrival period of $(\bar{A}_t)$. The connection between the two time scales is given precisely by the random increasing process $(\tau_u)$, defined next.

Let $\tau_k$ denote the time of the $k$th arrival, that is, $\tau_k = \min\{t : \bar{A}_t \geq k\}$. With $\tau_0 = 0$, the sequence $(\tau_k)_{k\geq 0}$ is a random walk, and the distribution of $\tau_1$ is exponential with mean $\alpha^{-1}$. Since the exponential distribution is infinitely divisible the process $(\tau_k)$ can be extended to a continuous-parameter process $(\tau_u)_{u\in\mathbb{R}_+}$ which is right-continuous and nondecreasing with probability one, and has stationary, independent increments (see [3, Sect. 14.4]. Of course the extension may require an enlargement of the basic probability space, and should be done so that $(\tau_u)$ is independent of $(\bar{D}_t)$). For each $u$, $\tau_u$ has a gamma distribution with Laplace transform

$$Ee^{-s\tau_u} = \left(\frac{\alpha}{s+\alpha}\right)^u. \tag{5.2}$$

The number of potential departures between the arrivals of customers $i$ and $i + 1$ at queue one is $D_{u_{i+1}} - D_{u_i}$ where $(D_u)_{u\geq 0}$ is defined by

$$D_u = \bar{D}_{\tau_u}. \tag{5.3}$$

Hence,

$$\tilde{N}_{i+1} = \max(0, \tilde{N}_i + 1 - (D_{u_{i+1}} - D_{u_i})). \tag{5.4}$$

Since the sequence $(u_i)$ marks the jump times of $(A_u)$, eq. (5.4) shows that $\tilde{N}_i$ is the number of customers just before the $i$th arrival in a queue with arrival process $(A_u)$ and potential service process $(D_u)$. Since the two processes $(\bar{D}_t)$ and $(\tau_u)$ are independent and each has stationary, independent increments, and since $(\tau_u)$ is nondecreasing with probability one, it easily follows that the process $(D_u)$ has stationary, independent increments. ($(D_u)$ is called $(\bar{D}_t)$ subordinated by $(\tau_u)$.) Since $(D_u)$ is also nondecreasing and integer-valued, it is thus a compound Poisson process. Therefore Theorem 1.1 can be applied to a queue which is governed by the processes $(A_u)$ and $(D_u)$. If $\lambda$ is replaced by $p$ in Theorem 1.1, then the conditions of Theorem 1.1 and Proposition 5.1 are identical. To complete the proof of Proposition 5.1 it remains only to identify the jump rate $\nu$ and jump size distribution $(q_k)$ of the process $(D_u)$ in order to show that eq. (5.1) for $x^*$ is the same as eq. (1.8).

Now, the generating function for $D_u$ is (using (5.2) and (5.3))

$$\phi_u(x) = Ex^{D_u} = EE[x^{D_u}|\tau_u] = Ee^{-(1-x)\mu\tau_u} = \left(\frac{\alpha}{(1-x)\mu+\alpha}\right)^u.$$

Taking the generating function of each side of eq. (1.2) and noting that the convolution on the right becomes multiplication yields that

$$\frac{d\phi_u}{du} = \phi_u(x)Q(x)$$

where $Q$ is defined by eq. (1.7). Hence, since $\phi_0(x) \equiv 1$,

$$Q(x) = \frac{d\phi_u}{du}(x)\bigg|_{u=0} = \log\left(\frac{\alpha}{(1-x)\mu+\alpha}\right). \tag{5.5}$$

By inversion, this implies that

$$q_k = \frac{1}{\nu k}\left(\frac{\mu}{\alpha+\mu}\right)^k \quad \text{and} \quad \nu = \log\left(1+\frac{\mu}{\alpha}\right).$$

Equation (5.5) shows that eqs. (5.1) and (1.8) are equivalent. Thus Proposition 5.1 is indeed implied by Theorem 1.1. □

*Appendix A*

The purpose of this appendix is to show that neither condition (1.11) nor condition (1.12) implies the other. Examples for which (1.12) holds but (1.11) does not are easily obtained by considering the special case $\tau_n = nX$ for all $n$ for some random variable $X$, for then (1.11) and (1.12) become $EX \leq \lambda^{-1}$ and $E[X^{-1}] \geq \lambda$, respectively. Thus in the remainder of this appendix we give an example for which (1.11) holds but (1.12) does not.

Let $T_1, T_2, \ldots$ be an increasing sequence of positive integers such that

$$T_1 + T_2 + \cdots + T_{i-1} + 1 \leq T_i \qquad \text{for} \quad i \geq 1.$$

Let $(U_{i,j}: i \geq 1, 1 \leq j \leq T_i)$ be an array of random variables such that for each $i$ fixed row $i$ of the array takes values in the set

$$\{(T_i, 0, 0, \ldots, 0), (0, T_i, 0, \ldots, 0), \ldots, (0, 0, \ldots, 0, T_i)\}$$

and is equal to any one element of the set with equal probability. Consider the arrival process such that $\tau_n$ is equal to one plus the sum of the first $n$ variables of the array, taken in lexicographic order. Now $EU_{i,j} = 1$ for each $i$, $j$ so that $E\tau_n = n + 1$ and thus $E\tau_n/n$ converges to one as $n$ tends to infinity.

On the other hand, since the first $T_i + \cdots + T_{i-1}$ customers all arrive by time $T_i$ and since the number of additional customers which arrive by time $T_i$ is uniformly distributed on the finite set $\{0, 1, \ldots, T_i - 1\}$, we have that

$$\begin{aligned} EA_{T_i} &= T_1 + T_2 + \cdots + T_{i-1} + \frac{1 + \cdots + T_i - 1}{T_i} \\ &= T_1 + T_2 + \cdots + T_{i-1} + \frac{T_i - 1}{2}. \end{aligned}$$

Thus, by choosing the sequence $(T_i)$ to increase very quickly we can arrange for $EA_{T_i}/T_i$ to converge to one half as $i$ tends to infinity and then

$$\liminf_{T\to\infty} \frac{EA_T}{T} = \frac{1}{2}.$$

Condition (1.11) is satisfied for $\lambda = 1$ but condition (1.12) is not.

## *Appendix B*

In this appendix the following proposition is proved.

PROPOSITION B1. *Suppose that $\Phi$ is a convex, nondecreasing function on $\mathbb{R}_+$ and that $\Psi$ is defined on $\mathbb{Z}_+$ by*

$$\Psi(n) = E\Phi(X_1 + X_2 + \cdots + X_n)$$

*where $X_1, X_2, \ldots$ are independent exponential random variables with parameter $\mu$. Then $\Psi$ is convex and nondecreasing on $\mathbb{Z}_+$.*

LEMMA B2. *Any convex, nondecreasing function $\Phi$ on $\mathbb{R}_+$ has a representation*

$$\Phi(x) = \Phi(0) + \int_0^\infty \Phi_c(x)\sigma(dc)$$

*for some positive measure $\sigma$ on $I\!R_+$ where $\Phi_c(x) = max(x - c, 0)$.*

PROOF OF LEMMA. Such a function $\Phi$ has a right-derivative $\Phi'_+(y)$ which is nonnegative, nondecreasing and right-continuous, such that [14, Corollary 24.4.1]

$$\Phi(x) = \Phi(0) + \int_0^x \Psi'_+(y)\,dy.$$

Thus, if $\sigma$ denotes the positive measure on $\mathbb{R}_+$ such that $\sigma(\{x{:}0 \le x \le y\}) = \Phi'_+(y)$, then by interchanging the order of integration,

$$\begin{aligned} \Phi(x) &= \Phi(0) + \int_{0-}^x \int_{0-}^y \sigma(dc)\,dy \\ &= \Phi(0) + \int_{0-}^x \left(\int_c^x dy\right)\sigma(dc) \end{aligned}$$

which yields the desired representation. □

PROOF OF PROPOSITION. With $\Psi_c(n) = E\Phi_c(X_1 + \cdots + X_n)$, direct computation yields that (assuming for simplicity that $\mu = 1$)

$$\Psi_c(n+1) - \Psi_c(n) = \int_c^\infty \frac{x^n}{n!} e^{-x}\,dx$$

and

$$\Psi_c(n+2) - 2\Psi_c(n+1) + \Psi_c(n) = \frac{c^{n+1}e^{-c}}{(n+1)!}$$

so that $\Psi_c$ is nondecreasing and convex for each fixed $c > 0$. By Lemma B2 and Fubini's Theorem which justifies changing the order of integration for nonnegative integrands,

$$\Psi(n) = \Phi(0) + E\int_0^\infty \Phi_c(X_1 + X_2 + \cdots + X_n)\sigma(dc)$$

$$= \Phi(0) + \int_0^\infty \Psi_c(n)\sigma(dc).$$

Since for each $c$, $\Psi_c$ is nondecreasing and convex, the same is true of $\Psi$. □

ACKNOWLEDGMENTS. I gratefully acknowledge Prof. A. Ephremides for introducing me to the subjects of this paper. I would also like to thank Mr. Tim van Loon for providing Figure 1, Prof. Sheldon Ross for suggesting reference [18], and Prof. P. Humblet for sending me his preprint [10].

REFERENCES

1. BERTSEKAS, D.P. Algorithms for nonlinear multicomodity network flow. *Int. Symp. on Systems Optimization and Analysis*, A. Bensoussan and J.L. Lions, Eds., Springer-Verlag, New York, 1979, pp 210–224.
2. BOROVKOV, A.A. *Stochastic Processes in Queueing Theory*. Springer-Verlag, New York, 1976.
3. BREIMAN, L. *Probability*. Addison-Wesley, Reading, Mass , 1968.
4. EPHREMIDES, A. Regularity and delay in queueing systems. In *Proc 18th Annual Conference on Communication, Control, and Computing*, Coordinated Science Laboratory, University of Illinois (Urbana, Ill., Oct. 1980), p. 339.
5. EPHREMIDES, A., VARAIYA, P., AND WALRAND, J. A simple dynamic routing problem. *IEEE Trans. Automatic Control 25*, 4 (Aug. 1980), 690–693.
6. FOSCHINI, G J., AND SALZ, J A basic dynamic routing problem and diffusion. *IEEE Trans. Commun. 26*, 3 (Mar. 1978), 320–327.
7. FRATTA, L., GERLA, M., AND KLEINROCK, L. The flow deviation method—an approach to store and forward communication network design. *Networks 3* (1973), 97–133.
8. FUCHS, L. A new proof of an inequality of Hardy-Littlewood-Polya. *Matematisk Tidsskritt B* (1974), 53–54.
9. GALLAGHER, R. A minimum delay routing algorithm using distributed computation. *IEEE Trans. Commun. 25*, 1 (Jan. 1977), 73–85
10. HUMBLET, P.A. Determinism minimizes waiting time in queues. Preprint. Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Mass., May 1982.
11. JACKSON, J.P. Networks of waiting lines. *Oper. Res. 5* (1957), 518–521.
12. KLEINROCK, L. *Queueing Systems—Volume 2: Computer Applications*. Wiley, New York, 1976.
13. NEUTS, M.F. *Matrix-Geometric Solutions in Stochastic Models—An Algorithmic Approach*. Johns Hopkins University Press, Baltimore, Md., 1981.
14. ROCKAFELLAR, R.T *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.
15. ROGOZIN, B.A. Some extremal problems in the theory of mass service. *Theor. Probab. Appl. 11*, 1 (1966), 144–151.
16. SEGALL, A. Optimal distributed routing for line-switched data networks. *IEEE Trans. Commun. 27*, 1 (Jan. 1979), 201–209.
17. TAKACS, L. *Introduction to the Theory of Queues*. Oxford University Press, New York, 1962.
18. VASICEK, O.A An inequality for the variance of waiting time under a general queueing discipline. *Oper. Res. 25*, 5 (Sept.-Oct. 1977), 879–884.
19. VAN LOON, T G. Application of phase-type Markov processes to multiple access and routing for packet communication. M.S. thesis, Dept. of Elec. Eng., University of Illinois, Urbana, Ill. (1981).
20. YUM, T.P The design and analysis of a semidynamic deterministic routing rule. *IEEE Trans. Commun. 29*, 4 (April 1981), 498–504.

RECEIVED JULY 1981; REVISED NOVEMBER 1982; ACCEPTED DECEMBER 1982