Logarithmic Asymptotics for Steady-State Tail Probabilities in a Single-Server Queue

Peter W. Glynn; Ward Whitt

Journal of Applied Probability, Vol. 31, Studies in Applied Probability (1994), 131-156.

Stable URL:
http://links.jstor.org/sici?sici=0021-9002%281994%2931%3C131%3ALAFSTP%3E2.0.CO%3B2-1

Journal of Applied Probability is currently published by Applied Probability Trust.

Logarithmic asymptotics for steady-state tail probabilities in a single-server queue

PETER W. GLYNN AND WARD WHITT

Abstract

We consider the standard single-server queue with unlimited waiting space and the first-in first-out service discipline, but without any explicit independence conditions on the interarrival and service times. We find conditions for the steady-state waiting-time distribution to have asymptotics of the form $x^{-1}\log \boldsymbol{P}(W > x) \to -\theta^*$ as $x \to \infty$ for $\theta^* > 0$. We require only stationarity of the basic sequence of service times minus interarrival times and a Gärtner–Ellis condition for the cumulant generating function of the associated partial sums, i.e. $n^{-1}\log \boldsymbol{E}\exp(\theta S_n) \to \psi(\theta)$ as $n \to \infty$, plus regularity conditions on the decay rate function ψ. The asymptotic decay rate θ^* is the root of the equation $\psi(\theta) = 0$. This result in turn implies a corresponding asymptotic result for the steady-state workload in a queue with general non-decreasing input. This asymptotic result covers the case of multiple independent sources, so that it provides additional theoretical support for a concept of effective bandwidths for admission control in multiclass queues based on asymptotic decay rates.

WAITING TIME; WORKLOAD, TAIL PROBABILITIES; ASYMPTOTICS; LARGE DEVIATIONS; CUMULANT GENERATING FUNCTION; COUNTING PROCESS; EFFECTIVE BANDWIDTHS; ADMISSION CONTROL; HIGH-SPEED NETWORKS; ASYNCHRONOUS TRANSFER MODE

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60K25

1. Introduction and summary

We are pleased to be able to contribute to this Festschrift in honor of Lajos Takács on his 70th birthday. In this paper we try to emulate Takács by seeking the essential mathematics underlying a probability problem of applied relevance. Like Takács (1962), (1963), (1967), we focus on the single-server queue.

In particular, we focus on asymptotics for the steady-state waiting time W and the steady-state workload L. We find general conditions under which

$$x^{-1}\log \boldsymbol{P}(W > x) \to -\theta^* \text{ as } x \to \infty \tag{1.1}$$

for $\theta^* > 0$, and similarly for L. We call the constant θ^* in (1.1) the *asymptotic decay rate*. The following elementary proposition helps put (1.1) in perspective. It is easily proved using integration by parts; see for instance p. 150 of Feller (1971).

Proposition 1. *For any random variable* Z *and positive constant* θ^*, *the following are equivalent*:

(i) $x^{-1}\log \boldsymbol{P}(Z > x) \to -\theta^*$ as $x \to \infty$;

(ii) $\sup\{\theta \geq 0 : E\exp(\theta Z) < \infty\} = \theta^*$;

(iii) *For all* $\epsilon > 0$, *there is an* $x_0 \equiv x_0(\epsilon)$ *such that*

$$\exp(-(\theta^* + \epsilon)x) \leq \boldsymbol{P}(Z > x) \leq \exp(-(\theta^* - \epsilon)x) \textit{ for all.} x > x_0.$$

There is currently great interest in asymptotics such as in (1.1) because of possible applications to the design and control of emerging high-speed communication networks. In particular, it has been recognized that asymptotic decay rate functions (defined below) that determine asymptotic decay rates such as θ^* in (1.1) may be used to create effective bandwidths for admission control and other network resource allocation problems; see Gibbens and Hunt (1991), Kelly (1991), Guerin et al. (1991), Chang (1993), Chang et al. (1992), Elwalid and Mitra (1993), Sohraby (1993), Whitt (1994) and references in these sources. Chang et al. (1992) also illustrates how the asymptotic decay rates may be used to speed up simulations. Our approach here is most closely related to the papers by Whitt (1994), Chang (1993) and Chang et al. (1992). In particular, the results here provide theoretical support for the procedures in Whitt (1994).

In many cases, a stronger limit than (1.1) holds, namely,

$$\exp(\theta^* x)\boldsymbol{P}(W > x) \to \alpha^* \text{ as } x \to \infty \tag{1.2}$$

for positive constants θ^* and α^*. Then we call α^* the *asymptotic constant*. It is easy to see that (1.2) implies (1.1) but not conversely. An $M/G/1$ queue for which (1.1) holds but (1.2) does *not* appears in Example 5 of Abate et al. (1994a). Then $\boldsymbol{P}(W > x) \sim \alpha x^{-3/2}\exp(-\theta^* x)$ as $x \to \infty$, where $f(x) \sim g(x)$ means that $f(x)/g(x) \to 1$. *In this paper we focus on the weaker form* (1.1). For work focusing on (1.2), see Abate et al. (1994a, b, c), Asmussen (1989), (1993), Asmussen and Perry (1992), Elwalid and Mitra (1993), (1994), Neuts (1986), Stern and Elwalid (1991), Tijms (1986), van Ommeren (1988) and references in these sources.

When (1.2) holds, a natural approximation for the tail probabilities is $\boldsymbol{P}(W > x) \approx \alpha^*\exp(-\theta^* x)$ for x not too small. Since the asymptotic constant α^* in (1.2) is often not easy to obtain, Abate et al. (1994a) suggest the simple approximation $\alpha^* \approx \theta^* \boldsymbol{E}W$. For some purposes, e.g. for percentiles, even $\alpha \approx 1$ is satisfactory. In many cases, $\alpha^* \approx 1$ produces a bound, i.e. $\boldsymbol{P}(W > x) \leq \exp(-\theta^* x)$; see p. 269 of Asmussen (1987) and Chang (1993). These exponential approximations can also be used with (1.1), even though (1.1) does not provide as much support as (1.2). However, Example 5 of Abate et al. (1994a) shows that the quality of the approximation provided by the asymptotics can deteriorate dramatically when (1.1) holds but (1.2) does not. Moreover, for the admission control problem, it is important to note that the quality of the approximations for the tail probabilities provided by the simple one-term exponential approximations also can deteriorate dramatically when the number of independent sources increases; see Choudhury et al. (1993), (1994).

In this first section, we present our main result and discuss its implications. We give proofs in Sections 2–8 and an example in Section 9. In Section 1.1 we state our main result for W; in Section 1.2 we discuss some implications and related results; in Section 1.3 we state our main results for L, which follow directly from the results for W by discretizing the processes; and in Section 1.4 we give sufficient conditions for W and L to have the same logarithmic asymptotics. This involves the logarithmic asymptotics of the time-stationary and customer-stationary (embedded-stationary or Palm-stationary) versions of the arrival process. In Section 1.5 we discuss logarithmic asymptotics for steady-state queue lengths.

1.1. *The main result*

Let $\{X_n : n \geq 1\}$ be a sequence of real-valued random variables and define the associated waiting-time sequence $\{W_n : n \geq 0\}$ recursively by letting $W_0 = 0$ and

$$W_{n+1} = [W_n + X_{n+1}]^+, \qquad n \geq 0, \tag{1.3}$$

where $[x]^+ = \max\{x, 0\}$. Let $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$, $n \geq 1$. Let $\Rightarrow$ denote convergence in distribution.

Theorem 1. *Let $\{X_n : n \geq 1\}$ be strictly stationary. If there exists a function ψ and positive constants θ^* and ϵ^* such that*

$$\text{(i)}\ \mathrm{n}^{-1}\log \boldsymbol{E}\exp(\theta \mathrm{S_n}) \to \psi(\theta) \text{ as n} \to \infty \text{ for } |\theta - \theta^*| < \epsilon^*, \tag{1.4}$$

$$\text{(ii) } \psi \text{ is finite in a neighbourhood of } \theta^*$$

$$\textit{and differentiable at } \theta^* \textit{ with } \psi(\theta^*) = 0 \textit{ and}$$

$$\psi'(\theta^*) > 0, \textit{ and} \tag{1.5}$$

$$\text{(iii) } \boldsymbol{E}\exp(\theta^* \mathrm{S_n}) < \infty \text{ for n} \geq 1, \tag{1.6}$$

then $W_n \Rightarrow W$ as $n \to \infty$ and (1.1) *holds.*

A significant feature of Theorem 1 is that there are no independence or Markov assumptions. Instead, we have condition (1.4) involving the asymptotic behavior of the cumulant generating functions of the partial sums S_n, as in the Gärtner (1977)–Ellis (1984) theorem of large deviations theory; see p. 14 of Bucklew (1990). (For a discussion of the connection to cumulants, see Choudhury and Whitt (1994).) Indeed, our proof of Theorem 1 follows large deviations theory, using exponential changes of measure. For additional background on large deviations theory, see Dembo and Zeitouni (1992) and Shwartz and Weiss (1993). Theorem 3.9 (ii) of Chang (1993), which was obtained independently and concurrently, is a result closely related to Theorem 1. The upper bound was obtained earlier in Chang (1992).

The conditions in Theorem 1 are very general, but they are not necessary, as we show in Example 1 in Section 9.

A (familiar) key step in proving Theorem 1 is representing W_n as the maximum of reverse-time partial sums; i.e.

$$W_n = S_n - \min_{0 \le k \le n} S_k = \max_{0 \le k \le n} \{S_n - S_k\}. \tag{1.7}$$

so that, when we extend $\{X_n\}$ to a doubly infinite stationary sequence $\{X_n : -\infty < n < \infty\}$, W_n is distributed as $\tilde{M}_n \equiv \max\{\tilde{S}_k : 0 \le k \le n\}$, where $\tilde{S}_0 = 0$ and $\tilde{S}_k = X_{-1} + \cdots + X_{-k}$. The conditions in Theorem 1 obviously apply to $\tilde{S}_k$ as well as S_k because $\boldsymbol{E}\exp(\theta\tilde{S}_k) = \boldsymbol{E}\exp(\theta S_k)$. Since the stationarity is required only for this step, we obtain Theorem 1 immediately from the following result for maxima of partial sums $M_n = \max\{S_k : 0 \le k \le n\}$, which does not require stationarity. We prove the following result in Section 2.

Theorem 2. *Let* $\{X_n : n \ge 1\}$ *be a sequence of real-valued random variables* (*not necessarily stationary or mutually independent*). *If there exists a function* ψ *and positive constants* θ^* *and* ϵ^* *such that* (1.4)–(1.6) *hold and*

$$\limsup_{n\to\infty} \boldsymbol{E}\exp(\theta X_n) < \infty \textit{ for } |\theta| < \epsilon^*, \tag{1.8}$$

then $M_n \to M$ *w.p.*1 *as* $n \to \infty$ *and*

$$x^{-1}\log \boldsymbol{P}(M > x) \to -\theta^* \textit{ as } x \to \infty. \tag{1.9}$$

Note that condition (1.8) in Theorem 2 is not needed if $\{X_n\}$ is stationary, because then (1.8) is implied by (1.6) in the case $n = 1$. For this, recall that $\boldsymbol{E}\exp(\theta_1 Z) < \infty$ when $\boldsymbol{E}\exp(\theta_2 Z) < \infty$ and $\theta_1 < \theta_2$ for any random variable Z by Hölder's inequality; see (21) on p. 47 of Chung (1974).

Also note that condition (1.6) is clearly necessary in Theorem 2, because $M \ge S_n$ for all n. Hence, $\boldsymbol{E}\exp(\theta^* M) = \infty$ if $\boldsymbol{E}\exp(\theta^* S_n) = \infty$ for any n.

We remark that we have also proved a version of Theorem 2 with condition (1.8) replaced by $\boldsymbol{E}\exp(\theta S_n) < \infty$ for $n \ge 1$ for some θ with $\theta > \theta^*$. This alternative condition might be preferred in Theorem 2, but it would require that we strengthen (1.6) in Theorem 1.

In Theorem 1 we have assumed that the basic sequence $\{X_n\}$ is stationary. However, this is not a great restriction because the focus is on the steady-state waiting time W. Given the distribution of W, it is usually possible to choose a stationary version of any given basic sequence $\{X_n\}$ such that $W_n \Rightarrow W$ as $n \to \infty$; see for example p. 13 of Borovkov (1976). Of course, the conditions in Theorem 1 apply to this stationary version. However, under regularity conditions, non-stationary versions and stationary versions of the basic sequence will couple so that the conditions for one enable us to verify the conditions for the other.

In other words, W typically does not depend on the initial part of the basic sequence $\{X_n\}$. In contrast, the maximum M in Theorem 2 clearly does depend on the entire sequence $\{X_n\}$. For a simple example, suppose that $\{X_n : n \geq 2\}$ is i.i.d. with a good distribution, but $\boldsymbol{P}(X_1 > x) \sim Ax^{-p}$. Then X_1 influences the tail behavior of S_n for all n and x, but not that of W.

An important role is played by the function ψ in Theorem 1; we call it the (*asymptotic*) *decay rate function*. It is significant that ψ is necessarily convex where it is finite, because $\log \boldsymbol{E}\exp(\theta Z)$ is convex where it is finite for any random variable Z, as can easily be seen by applying Hölder's inequality. It is important to distinguish the decay rate function ψ from the associated *large deviations rate function* $I(x)$, defined by

$$I(x) = \sup_{\theta}\{\theta x - \psi(\theta)\}; \tag{1.10}$$

see for example Chapter 1 of Bucklew (1990). The functions ϕ and I are intimately related. Indeed, they are convex conjugates of each other; see p. 183 of Bucklew (1990).

1.2. *Implications and related results*

The conditions of Theorems 1 and 2 are easy to check when the basic sequence $\{X_n\}$ is i.i.d. This special case includes the $GI/GI/1$ queue (with i.i.d. service times independent of i.i.d. interarrival times), for which it is possible to obtain the stronger result (1.2); see for example p. 269 of Asmussen (1987). For early results in this direction, see Smith (1953) and Theorems V.10.1 and VI.6.1 of Keilson (1965). Asmussen (1987), (1993) refers to the history in risk theory. In this $GI/GI/1$ case, Abate et al. (1993) have shown that it is also easy to compute the tail probabilities by numerical transform inversion, numerically integrating a contour integral representation for $\boldsymbol{E}\exp(-\theta W)$.

Corollary 1. *If* $\{X_n : n \geq 1\}$ *is i.i.d.,* $\boldsymbol{E}X < 0$,

$$\boldsymbol{E}\exp(\theta^* X) = 1 \tag{1.11}$$

and $\boldsymbol{E}\exp(\theta X) < \infty$ *for* $-\epsilon < \theta < \theta^* + \epsilon$ *for some* $\epsilon > 0$, *then the conditions of Theorems 1 and 2 hold with* $\psi(\theta) = \log \boldsymbol{E}\exp(\theta X)$, *so that* (1.1) *and* (1.9) *hold.*

Proof. Note that $n^{-1}\log \boldsymbol{E}\exp(\theta S_n) = \boldsymbol{E}\exp(\theta X_1) = \psi(\theta)$ when $\{X_n\}$ is i.i.d. Since $\psi(0) = 1$, $\psi'(0) = \boldsymbol{E}X < 0$ and ψ is convex, $\psi'(\theta^*) > 0$.

Thus, for the $GI/GI/1$ queue, it is easy to see what the decay rate function ψ is. For example, in the $M/M/1$ queue with service rate 1 and arrival rate ρ, $\psi(\theta) = -\log[(1-\theta)(1+\theta/\rho)]$; for the $D/M/1$ queue, $\psi(\theta) = -\log(1-\theta) - \theta/\rho$; and for the $M/D/1$ queue, $\psi(\theta) = \theta - \log(1+\theta/\rho)$.

It is worth pointing out that the logarithmic asymptotics in (1.1) tend to be robust. In general, weak convergence of distributions does not imply that large deviations asymptotics converge. However, in this context, weak convergence plus uniform integrability does imply that the cumulant generating function converges, and the logarithmic asymptotics here depend only on the location of the root (and not, for example, the slope at the root). We illustrate by stating a concrete result in the context of Corollary 1.

Corollary 2. *Let* $\{X_n^\gamma : n \geq 1\}$ *be i.i.d. for each* $\gamma > 0$ *where* $X_1^\gamma \Rightarrow X_1^0$ *as* $\gamma \to 0$ *and* $\boldsymbol{E}\exp(\theta X_1^\gamma) < M$ *for some* $\theta > \theta_0^*$ *and some finite* M, *for all* γ. *If* X_1^γ *satisfies the conditions of Corollary* 1 *for each* $\gamma \geq 0$, *then* (1.1) *holds for each* $\gamma \geq 0$ *and* $\theta_\gamma^* \to \theta_0^*$ *as* $\gamma \to 0$.

Proof. Since $X_1^\gamma \Rightarrow X_1^0$ as $\gamma \to 0$, $\exp(\theta X_1^\gamma) \Rightarrow \exp(\theta X_1^0)$ as $\gamma \to 0$. The uniform moment bound implies the uniform integrability needed to obtain $\boldsymbol{E}\exp(\theta X_1^\gamma) \to \boldsymbol{E}\exp(\theta X_1^0)$ as $\gamma \to 0$ for all $\theta < \theta_0^* + \epsilon$ for some ϵ.

In order to understand what the asymptotic decay rate θ^* in (1.1) primarily depends upon, and sometimes to compute θ^*, it is useful to consider heavy-traffic asymptotic expansions for θ^* based on expanding the function $\psi(\theta)$ in a Taylor series expansion about 0. Such heavy-traffic asymptotic expansions are established in Abate et al. (1994a), Abate and Whitt (1994) and Choudhury and Whitt (1994). Since $\log \boldsymbol{E}\exp(\theta S_n)$ is the cumulant generating function of S_n, the derivatives of $\psi(\theta)$ are the asymptotic cumulants of S_n. To illustrate, we establish the first term of the heavy-traffic expansion here. The first term coincides with the familiar decay rate associated with exponential heavy-traffic limits; see Kingman (1962), Iglehart and Whitt (1970) and Choudhury and Whitt (1994).

Corollary 3. *Consider a family of models indexed by* ρ, $0 < \rho < 1$. *Suppose that the assumptions of Theorem* 1 *hold for each* ρ *and*

(i) $\boldsymbol{E}X_n(\rho) = -(1-\rho)$,

(ii) $n^{-1}\operatorname{Var} S_n(\rho) \to \sigma^2$ *as* $n \to \infty$, $0 < \sigma^2 < \infty$, *and*

(iii) $n^{-1}\boldsymbol{E}(S_n(\rho) - (1-\rho)n)^3 \to \gamma$ *as* $n \to \infty, -\infty < \gamma < \infty$.

Then (1.1) *holds with*

$$\theta^*(\rho) = \frac{2(1-\rho)}{\sigma^2} + o(1-\rho) \ \textit{as}\ \rho \to 1.$$

Proof. Since $\log \boldsymbol{E}\exp(\theta S_n)$ is the cumulant generating function of S_n, we can apply Taylor's theorem to obtain

$$n^{-1}\log \boldsymbol{E}\exp(\theta S_n) = -\theta(1-\rho) + \frac{\theta^2 \operatorname{Var}(S_n)}{2n} + o(\theta^2) \text{ as } \theta \to 0$$

uniformly in ρ and η, using condition (iii) to get the uniformity in n; e.g., see (4′) on p. 268 of Chung (1974). Hence

$$\psi(\theta) = -\theta(1-\rho) + \frac{\theta^2\sigma^2}{2} + o(\theta^2) \text{ as } \theta \to 0$$

uniformly in ρ, so that the desired conclusion follows.

Another easy case is when the partial sums S_n are Gaussian (but possibly dependent) for all n. When S_n is Gaussian, Theorem 1 takes a very simple form. In particular, then (1.1) holds with θ^* in Corollary 3. The Gaussian assumption holds approximately in an $E_k/E_m/1$ queue for suitably large k and m. (As usual, E_k stands for Erlang of order k.) A direct Gaussian approximation has also been proposed and studied by Addie and Zuckerman (1993). This analysis provides additional justification for the heavy-traffic approximation, because it does not (at least directly) require a high traffic intensity.

Corollary 4. *Suppose that S_n is Gaussian with negative mean m_n and finite variance σ_n^2 for all $n \geq 1$. If $m_n/n \to m$ and $\sigma_n^2/n \to \sigma^2$ as $n \to \infty$, where $m < 0 < \sigma^2$, then the conditions of Theorem* 2 *hold with $\psi(\theta) = \theta m + \theta^2\sigma^2/2$, so that* (1.9) *holds with $\theta^* = -2m/\sigma^2 > 0$. If, in addition, the basic sequence $\{X_n\}$ is stationary, then the conditions of Theorem* 1 *hold, so that* (1.1) *holds.*

Proof. Recall that $\boldsymbol{E}\exp(\theta S_n) = \exp(\theta m_n + \theta^2\sigma^2/2)$ when S_n is Gaussian with mean m_n and variance σ_n^2.

In queueing theory, (1.3) is the familiar Lindley equation associated with a single-server queue with unlimited waiting room and the first-in first-out service discipline. Then $X_n = V_n - U_n$ where, for $n \geq 1$, V_n is the service time of customer n and U_n is the interarrival time between customers n and $n+1$. With this indexing, we begin with a first customer arriving at an empty system.

Another queueing model that leads to the representation $X_n = V_n - U_n$ is the queue length in a discrete-time single-server queue. Then we interpret V_n as the number of arrivals at epoch n and U_n as the number of *potential* departures at epoch n. For this representation to be valid, we usually require special Markov or deterministic assumptions in the service process, or 'autonomous service'; see p. 235 of Borovkov (1976). We use this below in Section 1.3. For the ATM networks it is often reasonable to assume deterministic service, so that this Lindley equation representation is indeed appropriate. For example, if there is at most one service completion at each epoch, then $U_n = 1$ for all n. This model variant is considered by Chang (1993), Sohraby (1993) and Chang et al. (1992). Chang (1993) also focuses on the Gärtner–Ellis condition in (1.4).

Given Theorems 1 and 2, we want to know when the conditions are satisfied. In the queueing context, the conditions can be expressed in terms of the two sequences $\{U_n : n \geq 1\}$ and $\{V_n : n \geq 1\}$ separately when the sequences $\{U_n\}$ and $\{V_n\}$ are independent. (However, note that such independence is not required in Theorem 1.)

To state the result, let $S_n^v = V_1 + \cdots + V_n$ and let $S_n^u = U_1 + \cdots + U_n$.

Proposition 2. *Suppose that $X_n = V_n - U_n$, where $\{V_n : n \geq 1\}$ and $\{U_n : n \geq 1\}$ are independent sequences of non-negative random variables* (*without independence or stationarity assumptions for each sequence*). *If there exist functions ψ_v and ψ_u*

and positive constants θ^*, ϵ^* *and* M *such that*

(1.12) (i) $n^{-1}\log \boldsymbol{E}\exp(\theta S_n^v) \to \psi_v(\theta)$ *as* $n \to \infty$ *for* $|\theta - \theta^*| < \epsilon^*$,

(1.13) (ii) ψ_v *is finite in a neighborhood of* θ^* *and differentiable at* θ^*,

(1.14) (iii) $\boldsymbol{E}\exp(\theta^* S_n^v) < \infty$ *for* $n \geq 1$,

(1.15) (iv) $\boldsymbol{E}\exp(\theta V_n) < M$ *for* $n \geq 1$ *and all* $|\theta| < \epsilon^*$,

(1.16) (v) $n^{-1}\log \boldsymbol{E}\exp(-\theta S_n^u) \to \psi_u(-\theta)$ *as* $n \to \infty$, *for* $|\theta - \theta^*| < \epsilon^*$,

(1.17) (vi) ψ_u *is finite in a neighborhood of* $-\theta^*$ *and differentiable at* $-\theta^*$,

(1.18) (vii) $\boldsymbol{E}\exp(-\theta U_n) < M$ *for* $n \geq 1$ *and all* $|\theta| < \epsilon^*$ *and*

(1.19) (viii) $\psi(\theta^*) = 0$ *and* $\psi'(\theta^*) > 0$ *for* $\psi(\theta) = \psi_v(\theta) + \psi_u(-\theta)$,

then $\{X_n\}$ *satisfies conditions* (1.4)–(1.6) *and* (1.8) *with decay rate function* ψ, *so that* (1.9) *holds. If, in addition,* $\{(U_n, V_n)\}$ *is stationary, then* $\{X_n\}$ *is stationary and* (1.1) *holds.*

Proof. By the independence,

$$\boldsymbol{E}\exp(\theta S_n) = \boldsymbol{E}\exp(\theta(S_n^v - S_n^u)) = \boldsymbol{E}\exp(\theta S_n^v)\boldsymbol{E}\exp(-\theta S_n^u),$$

so that

$$\log \boldsymbol{E}\exp(\theta S_n) = \log \boldsymbol{E}\exp(\theta S_n^v) + \log \boldsymbol{E}\exp(-\theta S_n^u).$$

Since $S_n^u \geq 0$, $\boldsymbol{E}\exp(-\theta S_n^u) \leq 1$. Similarly, $\boldsymbol{E}\exp(\theta X_n) = \boldsymbol{E}\exp(\theta V_n)\boldsymbol{E}\exp(-\theta U_n)$. Hence, it is clear that the assumed conditions here imply the conditions in Theorem 1.

Assuming that the arrival and service processes are independent, Proposition 2 shows that we can treat them separately, in the sense that the overall decay rate function ψ is the sum of the component decay rate functions, as indicated in (1.19). This separability is a basic feature of the heavy-traffic limits in Iglehart and Whitt (1970) and in the effective bandwidth approximations; see Elwalid and Mitra (1993), (1994), Stern and Elwalid (1991) and Whitt (1994). To obtain further results for these separate processes, it is useful to have a relation between the asymptotics for a counting process and the asymptotics for its inverse partial sum process. For this purpose, we apply a result from Glynn and Whitt (1994).

Let $\{T_n : n \geq 0\}$ be a non-decreasing sequence of random variables with $T_0 = 0$. We think of T_n as the arrival epoch of customer n in the queue; then $T_n = U_1 + \cdots + U_n$. Let $\{N(t) : t \geq 0\}$ be the associated counting process defined by

$$N(t) = \max\{n \geq 0 : T_n \leq t\}, \qquad t \geq 0. \tag{1.20}$$

The (familiar) key relation between T_n and $N(t)$ that we exploit is

$$\{N(t) \geq n\} = \{T_n \leq t\} \tag{1.21}$$

for all non-negative n and t.

A process $\{Z(t) : t \geq 0\}$ will be said to satisfy the *Gärtner–Ellis condition with decay rate function* ψ if

$$\lim_{t\to\infty} t^{-1} \log \boldsymbol{E} \exp^{\theta Z(t)} = \psi(\theta) \text{ for all } \theta \in \mathbb{R}. \tag{1.22}$$

For a discrete-time process, we let t run through the positive integers in (1.22).

The associated decay rate function ψ will be said to satisfy the *auxiliary large deviations* (*LD*) *regularity conditions* if (1.23)–(1.26) below hold:

$$\beta \equiv \inf\{\theta : \psi(\theta) = +\infty\} > 0, \tag{1.23}$$

$$\psi \text{ is differentiable everywhere in } (-\infty, \beta), \tag{1.24}$$

$$\lim_{\theta\uparrow\beta} \psi'(\theta) = +\infty, \tag{1.25}$$

and

$$\lim_{\theta\uparrow\beta} \psi(\theta) = \psi(\beta). \tag{1.26}$$

The conditions (1.22)–(1.26) are standard in the large deviations literature. In particular, under conditions (1.22)–(1.26), the process $\{Z(t) : t \geq 0\}$ satisfies the Gärtner (1977)–Ellis (1984) theorem, i.e. the *large deviations principle* holds for $\{Z(t) : t \geq 0\}$ with rate function I in (1.10); see pp. 42–50 of Dembo and Zeitouni (1992).

The following result is proved in Glynn and Whitt (1994). Let ψ^{-1} be the inverse function of ψ. Note that ψ is non-decreasing, and strictly increasing where it is finite. Hence, for x and y finite, $\psi^{-1}(y) = x$ if and only if $\psi(x) = y$.

Theorem 3. *If the counting process* $\{N(t) : t \geq 0\}$ *satisfies* (1.22)–(1.26), *then the inverse partial sum process* $\{T_n : n \geq 0\}$ *does too, with the possible exception of* (1.22) *for* $\theta = \beta_T$. *Similarly, if* $\{T_n : n \geq 0\}$ *satisfies* (1.22)–(1.26), *then* $\{N(t) : t \geq 0\}$ *does too, with the possible exception of* (1.22) *for* $\theta = \beta_N$. *In particular, then* (1.22) *holds for both processes, i.e.*

$$\lim_{t\to\infty} t^{-1} \log \boldsymbol{E} \exp(\theta N(t)) = \psi_N(\theta) \tag{1.27}$$

and

$$\lim_{n\to\infty} n^{-1} \log \boldsymbol{E} \exp(\theta T_n) = \psi_T(\theta) \tag{1.28}$$

both hold (*with the noted exceptions*) *and*

$$\psi_N(\theta) = -\psi_T^{-1}(-\theta), \tag{1.29}$$

where they are finite.

Thus, subject to regularity conditions, given the Gärtner–Ellis asymptotics for one of N or T, we obtain the Gärtner–Ellis asymptotics for the other directly and have the inverse relation (1.29). This parallels previous relations between other limits for N and T; see for example Iglehart and Whitt (1971), Section 7 of Whitt (1980), Theorems 3 and 6 of Glynn and Whitt (1988a) and Theorem 1 of Glynn and Whitt (1988b).

For example, we can apply Theorem 3 to obtain the Gärtner–Ellis limit (1.28) for the partial sums S_n^u from the Gärtner–Ellis limit (1.27) for the counting process $N(t)$ derived for batch Markovian arrival processes in Theorem 1 of Choudhury and Whitt (1994). Abate et al. (1994c) obtain (1.2) for $BMAP/GI/1$ queues, while the results here yield (1.1) for $BMAP/G/1$ queues, without requiring that the service times be i.i.d. Sufficient conditions for (1.22) in terms of embedded regenerative structure are also given in Theorem 7 of Glynn and Whitt (1994).

We now show that deterministic sequences provide upper bounds on θ^* when $\{U_n\}$ and $\{V_n\}$ are independent sequences; see Section 8 of Abate et al. (1994a) for related results. We use the queueing notation $G/G/1$ to refer to a general stationary sequence $\{(U_n, V_n)\}$ of interarrival times and service times.

Proposition 3. *Among $G/G/1$ models satisfying the assumptions of Proposition* 2, *the asymptotic decay rate θ^* is maximized* (*a*) *by deterministic service times among all stationary service-time sequences $\{V_n\}$ with given mean $\boldsymbol{E}V_n$, and* (*b*) *by deterministic interarrival times among all stationary interarrival-times sequences $\{U_n\}$ with given mean $\boldsymbol{E}U_n$.*

Proof. By Jensen's inequality, $\boldsymbol{E}\exp(\theta Z) \geq \exp(\theta \boldsymbol{E}Z)$ for any random variable Z, so that $\log \boldsymbol{E}\exp(\theta S_n^v) \geq \log\exp(\theta \boldsymbol{E}S_n^v) = n\theta \boldsymbol{E}V_1$ and $\log \boldsymbol{E}\exp(-\theta S_n^u) \geq \log\exp(\theta \boldsymbol{E}S_n^u) = n\theta \boldsymbol{E}U_1$. Hence, if ψ_v^{D} and ψ_u^{D} denote the decay rate functions in the deterministic cases, then $\psi_v(\theta) \geq \psi_v^{\mathrm{D}}(\theta)$ and $\psi_u(-\theta) \geq \psi_u^{\mathrm{D}}(-\theta)$ for all $\theta > 0$, so that the roots in (1.19) must be ordered as indicated.

More generally, we can establish stochastic comparisons between any two $G/G/1$ systems.

Proposition 4. *Consider two $G/G/1$ queues satisfying the assumptions of Theorem* 1. *If $\boldsymbol{E}\exp(\theta S_n^1) \leq \boldsymbol{E}\exp(\theta S_n^2)$ for all $\theta \geq 0$ and all n suitably large, then $\theta_2^* \leq \theta_1^*$.*

Proof. The condition implies that $\psi_1(\theta) \leq \psi_2(\theta)$ for all $\theta \geq 0$. Hence, the roots θ_i^* of $\psi_i(\theta) = 0$ must be ordered by $\theta_2^* \leq \theta_1^*$.

As in Whitt (1994), when $\{U_n\}$ and $\{V_n\}$ are independent, we can characterize the arrival and service decay rate functions $\psi_u(-\theta)$ and $\psi_v(\theta)$ from the asymptotic decay rates θ^* observed in $G/D/1$ and $D/G/1$ queues. To do this, we must consider all possible arrival rates ρ, $0 < \rho < 1$, so that the asymptotic decay rate θ^* becomes a *function* $\theta^*(\rho)$, $0 < \rho < 1$. Let $\psi_u(-\theta)$ refer to the case in which

$EU_n = 1$ and let the case of arrival rate ρ be obtained by considering interarrival times U_n/ρ for all n, i.e. simple time scaling.

Proposition 5. *For* $G/G/1$ *models satisfying the assumptions of Proposition* 2,

(i) *the arrival asymptotic decay rate function* $\psi_u(-\theta)$ *with arrival rate* 1 *is determined by the decay rate* $\theta^*(\rho)$ *in* $G/D/1$ *models with arrival rate* ρ, $0 < \rho < 1$, *i.e. by the equation*

$$\psi_u(-\theta^*(\rho)/\rho) + \theta^*(\rho) = 0, \qquad 0 < \rho < 1. \tag{1.30}$$

(ii) *The service asymptotic decay rate function* $\psi_v(\theta)$ *with service rate* 1 *is determined by the decay rate* $\theta^*(\rho)$ *in* $D/G/1$ *models with arrival rate* ρ, $0 < \rho < 1$, *i.e. by the equation*

$$\psi_v(\theta^*(\rho)) - \theta^*(\rho)/\rho = 0. \tag{1.31}$$

Proof. Note that $\psi_u(-\theta)$ is a decreasing convex function with $\psi_u'(0) = -1$. Hence, the values of $\psi_u(-\theta)$ for $\theta > 0$ are determined by the intersection with all lines through the origin with slopes less than -1. This is provided by (1.30), after making the change of variables $\theta(\rho) = \theta^*(\rho)/\rho$. Similarly, $\psi_v(\theta)$ is an increasing convex function with $\psi_v'(0) = 1$. Hence, the values of $\psi_v(\theta)$ for $\theta > 0$ are determined by the intersection with all lines through the origin with slope greater than $+1$. This is determined by (1.31).

Since Chang's (1993) model is equivalent to the $D/G/1$ special case, his equation $a^*(\theta) = c$ in (64) and the proof of Theorem 3.9 (ii) should coincide with (1.31), and it does.

1.3. *A continuous-time analog: the workload*

We can apply Theorems 1 and 2 to obtain corresponding results for continuous-time workload processes; we will only discuss the analog of Theorem 1. Paralleling (1.7), suppose that we have a continuous-time workload process $\{L(t) : t \geq 0\}$ defined in terms of a continuous-time net input process $\{Y(t) : t \geq 0\}$ by applying the usual reflection map, i.e.

$$L(t) = Y(t) - \inf\{Y(s) : 0 \leq s \leq t\}, \qquad t \geq 0, \tag{1.32}$$

with $L(0) = 0$. Moreover, let the net input process be defined in terms of a total input process $\{I(t) : t \geq 0\}$ with non-decreasing sample paths by

$$Y(t) = I(t) - t, \qquad t \geq 0. \tag{1.33}$$

In the $G/G/1$ queue, $I(t)$ represents the total work in service time to arrive in the interval $[0, t]$, i.e. the sum of all service times of all arrivals in $[0, t]$, but here $I(t)$ can be more general. For example, this formulation includes fluid models such as the Markov modulated fluid models in Elwalid and Mitra (1993) as a special case (without directly requiring the Markov assumption).

Paralleling Theorem 1, we will work with a version of $I(t)$ that has stationary increments. We prove the following result in Section 4.

Theorem 4. *Let the net input process* $\{Y(t) : t \geq 0\}$ *have stationary increments with* $\boldsymbol{E}Y(t) = (\rho - 1)t$ *where* $\rho < 1$. *If there exist a function* ψ *and positive constants* θ^* *and* ϵ^* *such that the analogs of* (1.4) *and* (1.6) *hold, i.e. if*

$$t^{-1} \log \boldsymbol{E} \exp(\theta Y(t)) \to \psi(\theta) \text{ as } t \to \infty \text{ for } |\theta - \theta^*| < \epsilon^* \tag{1.34}$$

and

$$\boldsymbol{E} \exp(\theta^* Y(t)) < \infty \text{ for all } t > 0, \tag{1.35}$$

and if (1.5) *holds for this* ψ, *then* $L(t) \Rightarrow L$ *as* $t \to \infty$ *and*

$$x^{-1} \log \boldsymbol{P}(L > x) \to -\theta^* \text{ as } x \to \infty. \tag{1.36}$$

Theorem 4 easily applies to superpositions of independent processes, as we now show.

Proposition 6. *Consider the workload process* $L(t)$ *in* (1.32) *and* (1.33) *with* $I(t) = I_1(t) + \cdots + I_n(t)$, *where* $I_1(t), \ldots, I_n(t)$ *are mutually independent nondecreasing processes each with stationary increments satisfying*

$$t^{-1} \log \boldsymbol{E} \exp(\theta I_i(t)) \to \psi_i(\theta) \text{ for } |\theta - \theta^*| < \epsilon^*$$

and

$$\boldsymbol{E} \exp(\theta^* I_i(t)) < \infty \text{ for all } t > 0 \text{ and } i.$$

If (1.5) *holds for*

$$\psi(\theta) = \psi_1(\theta) + \cdots + \psi_n(\theta) - \theta,$$

then the conditions of Theorem 4 *hold, so that* (1.36) *holds.*

Proof. By the independence,

$$\log \boldsymbol{E} \exp(\theta Y(t)) = \log \boldsymbol{E} \exp(\theta I_1(t)) + \cdots + \log \boldsymbol{E} \exp(\theta I_n(t)) - \theta t.$$

The following proposition treats the standard case in queueing, in which the total input $I(t)$ is the sum of all the service times of all arrivals in the interval $[0, t]$. We prove the following in Section 7.

Proposition 7. *Consider a total input process defined by*

$$I(t) = \sum_{i=1}^{A(t)} V_i, \qquad t \geq 0. \tag{1.37}$$

Suppose that $\{V_n\}$ *is independent of* $\{A(t)\}$,

$$n^{-1} \log \boldsymbol{E} \exp\left(\theta \sum_{i=1}^{n} V_i\right) \to \psi_v(\theta) \text{ as } n \to \infty \text{ for all } \theta \text{ in a neigborhood of } \hat{\theta}$$

and

$$t^{-1}\log \boldsymbol{E}\exp(\theta A(t)) \to \psi_A(\theta) \text{ as } t \to \infty \text{ for all } \theta \text{ in a neighborhood of } \psi(\hat{\theta}),$$

where ψ_A *is continuous at* $\psi_v(\hat{\theta})$. *Then*

$$t^{-1}\log \boldsymbol{E}\exp(\hat{\theta}I(t)) \to \psi_A(\psi_v(\hat{\theta})) \text{ as } t \to \infty.$$

Results related to Propositions 6 and 7 also appear in Chang (1993).

1.4. *Palm equivalence for the Gärtner–Ellis limits: relating W and L*

The asymptotics for W and L differ, in part, because W is based on the customer-stationary (embedded-stationary or Palm-stationary) sequence $\{U_n\}$ while L is based on the counting process $\{A(t)\}$ with stationary increments, which is associated with the time-stationary sequence, say $\{U_n^*\}$, connected by the Palm transformation, see for example Franken et al. (1981). However, we anticipate that we should have $\theta_W^* = \theta_L^*$. To establish that relation, we would like to have *Palm equivalence for the Gärtner–Ellis limits*, i.e. we would like to be able to say that $n^{-1}\log \boldsymbol{E}\exp(\theta S_n^u) \to \psi_u(\theta)$ as $n \to \infty$ if and only if $n^{-1}\log \boldsymbol{E}\exp(\theta S_n^{u*}) \to \psi_u^*(\theta)$ as $n \to \infty$ and $\psi_u = \psi_u^*$, where $S_n^{u*} = U_1^* + \cdots + U_n^*$, $n \geq 1$. We establish a weaker result here. We show that if both limits hold with the limit functions ψ_u and ψ_u^* satisfying regularity conditions, then $\psi_u = \psi_u^*$. We then use this property to provide conditions under which $\theta_L^* = \theta_W^*$.

We start by relating the asymptotics for L and W when the service times are i.i.d. As in Theorem 2 and Section 2 of Abate et al. (1994b), we apply the generalized Takács (1963) relation between W and L in a $G/GI/1$ queue (with i.i.d. service times that are independent of the arrival process); see (1.38) below and (4.5.9) on p. 129 of Franken et al. (1981).

Proposition 8. *In a* $G/GI/1$ *queue,* (1.1) *holds if and only if* (1.36) *holds and* $\theta_W^* = \theta_L^*$.

Proof. The generalized Takács relation yields

$$\boldsymbol{E}\exp(\theta L) = 1 - \rho + \rho \boldsymbol{E}\exp(\theta W)\boldsymbol{E}\exp(\theta V_{\mathrm{e}}), \tag{1.38}$$

where V_{e} has the stationary-excess or equilibrium-residual-life distribution associated with the service-time distribution. By Proposition 3 here, Theorem 10 of Abate et al. (1994a) and Lemma 1 of Abate et al. (1994b), $\boldsymbol{E}\exp(\theta V_{\mathrm{e}}) < \infty$ if $\boldsymbol{E}\exp(\theta W) < \infty$.

We now apply Proposition 8 to obtain a form of Palm equivalence for the Gärtner–Ellis limits. We prove the following result in Section 5.

Theorem 5. *Let* $A(t)$ *be a counting process associated with a non-deterministic time-stationary sequence* $\{U_n^*\}$ *and let* S_n^u *be the partial sums associated with the corresponding customer-stationary sequence* $\{U_n\}$. *Assume that the decay rate* *function associated with A, ψ_A, satisfies the auxiliary LD regularity conditions* (1.23)–(1.26) *with limit of support β_A in* (1.23). *Assume that*

$$\boldsymbol{E}\exp(\theta A(t)) < \infty \text{ for all } t > 0 \text{ and } \theta < \beta_A \tag{1.39}$$

and

$$t^{-1}\log \boldsymbol{E}\exp(\theta A(t)) \to \psi_A(\theta) \text{ as } t \to \infty \text{ for } \theta < \beta_A. \tag{1.40}$$

Assume that

$$n^{-1}\log \boldsymbol{E}\exp(-\theta S_n^u) \to \psi_u(-\theta) \text{ as } n \to \infty \text{ for all } \theta > 0, \tag{1.41}$$

where $\psi_u(-\theta)$ is finite and differentiable for all $\theta > 0$. Then

$$\psi_u(-\theta) = \psi_u^*(-\theta) = -\psi_A^{-1}(\theta) \text{ for all } \theta > 0. \tag{1.42}$$

We now relate the logarithmic asymptotics for W and L in a general $G/G/1$ queue when the arrival and service processes are independent (but the service times need not be i.i.d.). We prove the following result in Section 6.

Theorem 6. *Consider a $G/G/1$ queue in which the service times $\{V_n\}$ are stationary and independent of the arrival process. Let the arrival process satisfy the assumptions of Theorem* 5. *Let the service decay rate function ψ_v satisfy the auxiliary LD regularity conditions* (1.23)–(1.26) *with limit of support β_v in* (1.23). *Assume that*

$$\boldsymbol{E}\exp(\theta S_n^v) < \infty \text{ for all } n \geq 1 \text{ and } \theta < \beta_v \tag{1.43}$$

and

$$n^{-1}\log \boldsymbol{E}\exp(\theta S_n^v) \to \psi_v(\theta) \text{ as } n \to \infty \text{ for } \theta < \beta_v. \tag{1.44}$$

Then (1.1) *and* (1.36) *both hold with $\theta_W^*(\rho) = \theta_L^*(\rho)$ for each ρ, $0 < \rho < 1$.*

1.5. *Queue lengths*

In this section we discuss the logarithmic asymptotics for the steady-state queue length (number in system). Let Q and Q^a be the steady-state queue length at an arbitrary time and at an arrival epoch, respectively, which we assume are well defined. As in Section 1.3, let $G/GI/1$ mean i.i.d. service times that are independent of general stationary interarrival times. We prove the following in Section 8.

Proposition 9. *In the $G/GI/1$ queue,* (1.1) *holds if and only if the analogs of* (1.1) *hold for Q and Q^a, in which case*

$$\theta_Q^* = \theta_{Q^a}^* = \log \boldsymbol{E}\exp(\theta_W^* V_1) = \psi_v(\theta_W^*). \tag{1.45}$$

2. Proof of Theorem 2

In this section we prove Theorem 2. For this purpose, we perform a change of measure for each n. In particular, for each $n \geq 1$, let P_n^* be the probability measure on $\mathbb{R}^n$ defined by

$$\begin{aligned} P_n^*(dx_1, \ldots, dx_n) &= \frac{\exp(\theta^* S_n)}{\boldsymbol{E}\exp(\theta^* S_n)} \boldsymbol{P}(X_1 \in dx_1, \ldots, X_n \in dx_n) \\ &= \exp(\theta^* S_n - \psi_n(\theta^*))\boldsymbol{P}(X_1 \in dx_1, \ldots, X_n \in dx_n), \end{aligned} \tag{2.1}$$

where $\psi_n(\theta) = \log \boldsymbol{E}\exp(\theta S_n)$ and $\psi_n(\theta^*) < \infty$ for $n \geq 1$ by (1.6).

We base our proof on the following strengthened form of the weak law of large numbers. This is closely related to claim 1 on p. 17 of Bucklew (1990) in his proof of the Gärtner–Ellis theorem. However, we only make assumptions locally around θ^*, whereas Bucklew's assumptions are more global. We will need the cases $k = 0$ and $k = 1$ in our proof of Theorem 2. We prove Theorem 7 in Section 3.

Theorem 7. *Let k be a fixed non-negative integer. Under the conditions of Theorem* 2 (*excluding* (1.8) *if* $k = 0$), *for each* $\epsilon > 0$ *there exists* n_0 *and* $\eta \equiv \eta(\epsilon) \in [0, 1)$ *such that*

$$P_n^*\left(\left|\frac{S_{n-k}}{n} - \nu\right| > \epsilon\right) \leq \eta^n \textit{ for } n \geq n_0. \tag{2.2}$$

Since M_n is non-decreasing, $M_n \to M$ w.p.1. The desired result (1.9) implies that M must be proper. Since $\boldsymbol{P}(M > x) = \boldsymbol{P}(T(x) < \infty)$, where

$$T(x) = \inf\{n \geq 0 : S_n > x\}, \tag{2.3}$$

it suffices to show that

$$x^{-1} \log \boldsymbol{P}(T(x) < \infty) \to -\theta^* \text{ as } x \to \infty. \tag{2.4}$$

Let $\lfloor x \rfloor$ be the greatest integer less than or equal to x and let $\lceil x \rceil$ be the least integer greater than or equal to x. Now, for ϵ and ν given, and any x and $n(\epsilon)$,

$$\begin{aligned} \boldsymbol{P}(T(x) < \infty) &= \sum_{j=1}^{\infty} \boldsymbol{P}(T(x) = j) \\ &\leq \sum_{j=1}^{n(\epsilon)} \boldsymbol{P}(T(x) = j) + \sum_{j=n(\epsilon)+1}^{\lfloor x(1-\epsilon)/\nu \rfloor} \boldsymbol{P}(T(x) = j) \\ &\quad + \sum_{j=\lceil x(1-\epsilon)/\nu \rceil}^{\lfloor x(1+\epsilon)/\nu \rfloor} \boldsymbol{P}(T(x) = j) + \sum_{j=\lceil x(1+\epsilon)/\nu \rceil}^{\infty} \boldsymbol{P}(T(x) = j). \end{aligned} \tag{2.5}$$

Given ϵ, we choose $n(\epsilon)$ in (2.5) so that for all $n \geq n(\epsilon)$ we simultaneously have

$$|n^{-1}\psi_n(\theta^*)| < \min\left\{-\frac{\log\eta}{2}, \epsilon\right\} \tag{2.6}$$

and

$$P_n^*\left(|n^{-1}S_{n-k} - \nu| > \frac{\epsilon\nu}{1+\epsilon}\right) \leq \eta^n \tag{2.7}$$

for $k = 0$ and 1 for some η with $0 \leq \eta < 1$. This is possible because of assumption (1.4) and Theorem 7.

For the first term in (2.5),

$$\begin{aligned} P(T(x) = j) &\leq P(S_j > x) \\ &\leq E_j^*[\exp(-\theta^* S_j + \psi_j(\theta^*)); S_j > x] \\ &\leq \exp(-\theta^* x) E_j^*[\exp(\psi_j(\theta^*)); S_j > x] \\ &\leq \exp(-\theta^* x)\exp(\psi_j(\theta^*)). \end{aligned} \tag{2.8}$$

We use (1.6) to ensure that (2.8) is finite.

For the second term in (2.5), note that (starting with the reasoning in (2.8))

$$P(T(x) = j) \leq \exp(-\theta^* x) E_j^*[\exp(\psi_j(\theta^*)); S_j > x],$$

where

$$\begin{aligned} E_j^*[\exp(\phi_j(\theta^*); S_j > x] &\leq \exp\left(j\left(\frac{-\log\eta}{2}\right)\right) P_j^*(S_j > x) \\ &\leq \exp\left(j\frac{(-\log\eta)}{2}\right) P_j^*\left(S_j > \frac{\nu j}{1-\epsilon}\right) \\ &\leq \exp\left(j\frac{(-\log\eta)}{2}\right) P_j^*\left(j^{-1}S_j - \nu > \frac{\nu}{1-\epsilon} - \nu\right) \\ &\leq \exp\left(j\frac{(-\log\eta)}{2}\right) P_j^*\left(|j^{-1}S_j - \nu| > \frac{\epsilon\nu}{1+\epsilon}\right) \\ &\leq \exp\left(j\left(\frac{-\log\eta}{2}\right)\right)\eta^j \leq \exp\left(j\frac{\log\eta}{2}\right). \end{aligned} \tag{2.9}$$

Hence,

$$\sum_{j=n(\epsilon)+1}^{\lceil x(1-\epsilon)/\nu\rceil} P(T(x) = j) \leq \exp(-\theta^* x)\sum_{j=1}^{\infty}\eta^{j/2} \leq \exp(-\theta^* x)(1-\sqrt{\eta})^{-1}. \tag{2.10}$$

For the third term in (2.5),

$$P(T(x) = j) \leq \exp(-\theta^* x) E_j^*[\exp\psi_j(\theta^*); S_j > x], \tag{2.11}$$

where

$$\text{(2.12)} \quad \boldsymbol{E}_j^*[\exp(\psi_j(\theta^*)); S_j > x] \le \exp(\psi_j(\theta^*)) \le \exp(j\epsilon) \le \exp(\epsilon x(1+\epsilon)/\nu).$$

For the fourth term in (2.5),

$$\begin{aligned} \boldsymbol{P}(T(x)=j) &\le \boldsymbol{P}(S_{j-1} \le x, S_j > x) \\ &\le \boldsymbol{E}_j^*[\exp(-\theta^* S_j + \psi_j(\theta^*)); S_{j-1} \le x, S_j > x] \\ &\le \exp(-\theta^* x)\exp(\psi_j(\theta^*))\boldsymbol{P}_j^*(S_{j-1} \le x), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{P}_j^*(S_{j-1} \le x) \le \boldsymbol{P}_j\left(S_{j-1} \le \frac{\nu j}{1+\epsilon}\right) &= \boldsymbol{P}_j^*\left(\frac{S_{j-1}}{j} \le \frac{\nu}{1+\epsilon}\right) \\ &\le \boldsymbol{P}_j^*\left(\left|\frac{S_{j-1}}{j} - \nu\right| > \frac{\epsilon\nu}{1+\epsilon}\right) \le \eta^j \end{aligned}$$

by Theorem 5 with $k = 1$. Since $\psi_j(\theta^*) \le -(\log\eta)/2$ by (2.6) for j in this sum,

$$\text{(2.13)} \quad \sum_{j=\lfloor x(1+\epsilon)/\nu \rfloor}^{\infty} \boldsymbol{P}(T(x)=j) \le \exp(-\theta^* x)\sum_{j=0}^{\infty}\eta^{j/2} \le \exp(-\theta^* x)(1-\sqrt{\eta})^{-1}.$$

Combining (2.5), (2.8), (2.10), (2.12) and (2.13), we obtain

$$\begin{aligned} \boldsymbol{P}(T(x) < \infty) \le \exp(-\theta^* x)\Bigg\{ &\sum_{j=1}^{n(\epsilon)} \exp(\psi_j(\theta^*)) \\ &+(1-\sqrt{\eta})^{-1} + \left(\frac{2\epsilon x}{\nu}+1\right)\exp(\epsilon x(1+\epsilon)/\nu) + (1-\sqrt{\eta})^{-1}\Bigg\}. \end{aligned}$$

Hence, using condition (1.6),

$$\limsup_{x\to\infty} x^{-1}\log \boldsymbol{P}(T(x) < \infty) \le -\theta^* + \epsilon(1+\epsilon)/\nu.$$

Since ϵ was arbitrary,

$$\text{(2.14)} \quad \limsup_{x\to\infty} x^{-1}\log \boldsymbol{P}(T(x) < \infty) \le -\theta^*.$$

We now establish the lower bound. For this purpose, let $m(\epsilon) = \lceil x(1+\epsilon)/\nu \rceil$. Then

$$\begin{aligned} \boldsymbol{P}(T(x) < \infty) &\ge \boldsymbol{P}(S_{m(\epsilon)} > x) \\ &\ge \boldsymbol{E}_{m(\epsilon)}^*[\exp(-\theta^* S_{m(\epsilon)} + \psi_{m(\epsilon)}(\theta^*)); S_{m(\epsilon)} > x] \\ &\ge \boldsymbol{E}_{m(\epsilon)}^*\left[\exp(-\theta^* S_{m(\epsilon)} + \psi_{m(\epsilon)}(\theta^*)); S_{m(\epsilon)} > \frac{m(\epsilon)\nu}{1+\epsilon}\right] \end{aligned}$$

$$\geq E^*_{m(\epsilon)}\left[\exp(-\theta^* S_{m(\epsilon)} + \psi_{m(\epsilon)}(\theta^*)); m(\epsilon)^{-1}|S_{m(\epsilon)} - \nu| < \frac{\epsilon\nu}{1+\epsilon}\right]$$

$$\geq E^*_{m(\epsilon)}\left[\exp(-\theta^*\nu\frac{(1+2\epsilon)}{1+\epsilon}m(\epsilon)) + \psi_{m(\epsilon)}(\theta^*); |m(\epsilon)^{-1}S_{m(\epsilon)} - \nu| < \frac{\epsilon\nu}{1+\epsilon}\right]$$

$$\geq \exp\left(-\theta^*\nu\frac{(1+2\epsilon)}{1+\epsilon}m(\epsilon) + \psi_{m(\epsilon)}(\theta^*)\right)P^*_{m(\epsilon)}\left(|m(\epsilon)^{-1}S_{m(\epsilon)} - \nu| < \frac{\epsilon\nu}{1+\epsilon}\right).$$

Since

$$P^*_{m(\epsilon)}\left(|m(\epsilon)^{-1}S_{m(\epsilon)} - \nu| < \frac{\epsilon\nu}{1+\epsilon}\right) \to 1 \text{ as } x \to \infty$$

by Theorem 5,

$$\liminf_{x\to\infty} x^{-1}\log P(T(x) < \infty) \geq \liminf_{x\to\infty}\left\{-\theta^*\nu\frac{(1+2\epsilon)}{1+\epsilon}\frac{m(\epsilon)}{x} - \frac{\epsilon m(\epsilon)}{x}\right\}$$

$$\geq -\theta^*(1+2\epsilon) - \epsilon(1+\epsilon)/\nu.$$

Since ϵ was arbitrary, we conclude that

$$\liminf_{x\to\infty} x^{-1}\log P(T(x) < \infty) \geq -\theta^*. \tag{2.15}$$

Combining (2.14) and (2.15) completes the proof.

3. Proof of Theorem 7

As before, let $\psi_n(\theta) = \log E\exp(\theta S_n)$ and recall that $n^{-1}\psi_n(\theta) \to \psi(\theta)$ as $n \to \infty$ for $|\theta - \theta^*| < \epsilon^*$ where θ^* and ϵ^* are as assumed. We start with the case $k = 0$. Then, for each θ, $0 < \theta < \epsilon^*$,

$$P^*_n(n^{-1}S_n > \nu + \epsilon) \leq \exp(-\theta n(\nu+\epsilon))E^*_n\exp(\theta S_n),$$

where

$$\exp(-\theta n(\nu+\epsilon))E^*_n\exp(\theta S_n) = \exp(-\theta n(\nu+\epsilon))$$
$$\times \int \exp(\theta x)\exp(\theta^* x - \psi_n(\theta^*))P(S_n \in dx)$$
$$= \exp(-\theta n(\nu+\epsilon))\exp(\psi_n(\theta+\theta^*) - \psi_n(\theta^*)).$$

We choose n_0 in Theorem 7 suitably large so that $\psi_n(\theta+\theta^*) < \infty$, which is possible by (1.4) and (1.5). We use the fact that $E\exp(\theta_1 Z) < \infty$ when $E\exp(\theta_2 Z) < \infty$ and $0 < \theta_1 < \theta_2$, hence

$$\limsup_{n\to\infty} n^{-1}\log P^*_n(n^{-1}S_n > \nu + \epsilon) \leq \psi(\theta^* + \theta) - \psi(\theta^*) - \theta(\nu + \epsilon). \tag{3.1}$$

However, by Taylor's theorem,

$$\begin{aligned}\psi(\theta+\theta^*)-\psi(\theta^*)-\theta(\nu+\epsilon)&=\psi'(\theta^*)\theta+o(\theta)-\theta(\nu+\epsilon)\text{ as }\theta\to 0\\&=-\theta\epsilon+o(\theta)\text{ as }\theta\to 0.\end{aligned}$$

Hence, we can choose θ_1 with $0<\theta_1<\epsilon^*$ so that

$$\psi(\theta^*+\theta_1)-\psi(\theta^*)-\theta_1(\nu+\epsilon)<-\theta_1\epsilon/2$$

and

$$\lim_{n\to\infty}\sup n^{-1}\log \boldsymbol{P}_n^*(n^{-1}S_n>\nu+\epsilon)\leq-\theta_1\epsilon/2,$$

which establishes one half of (2.2).

On the other hand, for $0<\theta<\epsilon^*$,

$$\begin{aligned}\boldsymbol{P}_n^*(n^{-1}S_n<\nu-\epsilon)&=\boldsymbol{P}_n^*(-\theta S_n>-\theta n(\nu-\epsilon))\\&\leq\exp(\theta n(\nu-\epsilon))\boldsymbol{E}_n^*\exp(-\theta S_n),\end{aligned}$$

where

$$\begin{aligned}\exp(\theta n(\nu-\epsilon))\boldsymbol{E}_n^*\exp(-\theta S_n)&=\exp(\theta n(\nu-\epsilon))\int\exp(-\theta x)\boldsymbol{P}_n^*(S_n\in dx)\\&=\exp(\theta n(\nu-\epsilon))\exp(\psi_n(\theta^*-\theta)-\psi_n(\theta^*)).\end{aligned}$$

Hence,

$$\lim_{n\to\infty}\sup \boldsymbol{P}_n^*(n^{-1}S_n<\nu-\epsilon)\leq\theta(\nu-\epsilon)+\psi(\theta^*-\theta)-\psi(\theta^*).\tag{3.2}$$

Then, as before,

$$\psi(\theta^*-\theta)-\psi(\theta^*)-\theta(\nu-\epsilon)=-\theta\epsilon+o(\theta)\text{ as }\theta\to 0$$

so that we can choose θ_2 with $0<\theta_2<\epsilon^*$ so that

$$\psi(\theta^*-\theta_2)-\psi(\theta^*)+\theta_2(\nu-\epsilon)\leq-\theta_2\epsilon/2$$

and

$$\lim_{n\to\infty}\sup n^{-1}\log \boldsymbol{P}_n^*(n^{-1}S_n<\nu-\epsilon)\leq-\theta_2\epsilon/2,$$

which completes the proof for $k=0$.

For $k\geq 1$, we first note that $\boldsymbol{E}\exp(\theta(S_n-S_{n-k}))<\infty$ for all $|\theta|<\delta$ for some $\delta>0$ if condition (1.8) holds. To see this, apply the Cauchy–Schwarz inequality k times to obtain

$$\begin{aligned}&\boldsymbol{E}\exp(\theta(S_n-S_{n-k}))\leq(\boldsymbol{E}\exp(\theta X_n)^2)^{1/2}\boldsymbol{E}(\exp(\theta(S_{n-1}-S_{n-k}))^2)^{1/2}\\&\leq(\boldsymbol{E}\exp(2\theta X_n))^{1/2}\boldsymbol{E}(\exp(2\theta(S_{n-1}-S_{n-k})))^{1/2}\\&\leq(\boldsymbol{E}\exp(2\theta X_n))^{1/2}(\boldsymbol{E}\exp(4\theta X_{n-1}))^{1/4}\ldots(\boldsymbol{E}\exp(2^{k+1}\theta X_{n-k+1}))^{2^{-(k+1)}}.\end{aligned}$$

We choose n_0 suitably large so that, for some finite M,

$$(\boldsymbol{E}\exp(2\theta X_n))^{1/2}\boldsymbol{E}\exp(4\theta X_{n-1}))^{1/4}\dots\boldsymbol{E}\exp(2^{k+1}\theta X_{n-k+1}))^{2^{-(k+1)}} < M$$

for all $n \geq n_0$, which is possible by assumption (1.8).

For $k \geq 1$, we then have

$$\begin{aligned}
\boldsymbol{P}_n^*(n^{-1}S_{n-k} > \nu + \epsilon) &\leq \exp(-\theta n(\nu+\epsilon))\boldsymbol{E}_n^*\exp(\theta S_{n-k}) \\
&\leq \exp(-\theta n(\nu+\epsilon))\int \exp\left(\theta\sum_{i=1}^{n-k} x_i\right)\exp\left(\theta^*\sum_{i=1}^{n} x_i - \psi_n(\theta^*)\right) \\
&\qquad \times \boldsymbol{P}(X_1 \in dx_1, \dots, X_n \in dx_n) \\
&\leq \exp(-\theta n(\nu+\epsilon))\boldsymbol{E}[\exp((\theta+\theta^*)S_n - \theta(S_n - S_{n-k}) - \psi_n(\theta^*))] \\
&\leq \exp(-\theta n(\nu+\epsilon))(\boldsymbol{E}\exp(p(\theta+\theta^*)S_n))^{1/p}(\boldsymbol{E}\exp(-q\theta(S_n - S_{n-k})))^{1/q}\exp(-\psi_n(\theta^*))
\end{aligned}$$

for positive p and q with $p^{-1} + q^{-1} = 1$ by Hölder's inequality. We choose p sufficiently close to 1 and θ sufficiently small so that $p(\theta + \theta^*)$ is within the required neighborhood of θ^* and $q\theta < \delta$, so that $\boldsymbol{E}\exp(-q\theta(S_n - S_{n-k}))^{1/q}$ is bounded for $n \geq n_0$. Hence,

$$\lim_{n\to\infty}\sup n^{-1}\log \boldsymbol{P}_n^*(n^{-1}S_{n-k} > \nu + \epsilon) \leq -\theta(\nu+\epsilon) + \psi(p(\theta+\theta^*))^{1/p} - \psi(\theta^*). \tag{3.3}$$

Since p was arbitrary, we can let $p \to 1$ in (3.3) to obtain the analog of (3.1) with S_{n-k} instead of S_n.

Similarly,

$$\begin{aligned}
\boldsymbol{P}_n^*(n^{-1}S_{n-k} < \nu - \epsilon) &\leq \exp(\theta n(\nu-\epsilon))\boldsymbol{E}_n^*\exp(-\theta S_{n-k}) \\
&\leq \exp(\theta n(\nu-\epsilon))\int \exp\left(-\theta\sum_{i=1}^{n-k} x_i\right)\exp\left(\theta^*\sum_{i=1}^{n} x_i - \psi_n(\theta^*)\right) \\
&\qquad \times \boldsymbol{P}(X_1 \in dx_1, \dots, X_n \in dx_n) \\
&\leq \exp(\theta n(\nu-\epsilon))\boldsymbol{E}[\exp(\theta^* - \theta)S_n + \theta(S_n - S_{n-k}) - \psi_n(\theta^*)] \\
&\leq \exp(\theta n(\nu-\epsilon))(\boldsymbol{E}\exp(p(\theta^* - \theta)S_n))^{1/p}(\boldsymbol{E}\exp(q\theta(S_n - S_{n-k})))^{1/q}\exp(-\psi_n(\theta^*))
\end{aligned}$$

for positive p and q with $p^{-1} + q^{-1} = 1$ by Hölder's inequality. Reasoning as in (3.3), we obtain

$$\lim_{n\to\infty}\sup n^{-1}\log \boldsymbol{P}_n^*(n^{-1}S_{n-k} < \nu - \epsilon) \leq \theta(\nu-\epsilon) + \psi(p(\theta^* - \theta))^{1/p} - \psi(\theta^*). \tag{3.4}$$

Letting $p \to 1$ in (3.4) we obtain the analog of (3.2) with S_{n-k} instead of S_n. The rest of the proof is the same as for $k = 0$.

4. Proof of Theorem 4

We construct discrete-time processes satisfying the conditions of Theorem 1 that suitably approximate the continuous-time processes. In particular, for any $\delta > 0$, we construct a discrete-time waiting-time process $\{W_n^\delta\}$ by defining service times V_n^δ and interarrival times U_n^δ via

$$V_n^\delta = I((n+1)\delta) - I(n\delta) \quad \text{and} \quad U_n^\delta = \delta, \quad n \geq 1. \tag{4.1}$$

Since $\boldsymbol{E}Y(t) = \rho t$ for $\rho < 1$, $\boldsymbol{E}V_n^\delta < \boldsymbol{E}U_n^\delta$. We initialize by setting $W_0^\delta = L(0)$. Then, by induction, we have

$$W_n^\delta \leq L(n\delta) \leq W_n^\delta + \delta, \qquad n \geq 0. \tag{4.2}$$

Since $Y(t)$ has stationary increments, $L(t)$ is distributed the same as $\sup\{Y(s) : 0 \leq s \leq t\}$. Since this supremum is non-decreasing, $L(t) \Rightarrow L$ as $t \to \infty$. Since $W_n^\delta \Rightarrow W^\delta$ and $L(n\delta) \Rightarrow L$ as $n \to \infty$, we have

$$\boldsymbol{E}\exp(\theta W^\delta) \leq \boldsymbol{E}\exp(\theta L) \leq \exp(\theta\delta)\boldsymbol{E}\exp(\theta W^\delta). \tag{4.3}$$

From (4.3), we see that $\boldsymbol{E}\exp(\theta L) < \infty$ if and only if $\boldsymbol{E}\exp(\theta W^\delta) < \infty$.

Hence, it suffices to show that $\sup\{\theta : \boldsymbol{E}\exp(\theta W^\delta) < \infty\} = \theta^*$. For this purpose, let $S_n^\delta = V_0^\delta + \cdots + V_{n-1}^\delta - n\delta$. Then

$$\begin{aligned} \boldsymbol{E}\exp(\theta(S^{\delta\lfloor t/\delta\rfloor} - \delta)) &= \boldsymbol{E}\exp(\theta(Y(\delta\lfloor t/\delta\rfloor) - \delta)) \leq \boldsymbol{E}\exp(\theta Y(t)) \\ &\leq \boldsymbol{E}\exp(\theta(Y(\delta\lceil t/\delta\rceil) + \delta)) \leq \boldsymbol{E}\exp(\theta(S^{\delta\lceil t/\delta\rceil} + \delta)). \end{aligned} \tag{4.4}$$

Therefore,

$$\begin{aligned} \limsup_{t\to\infty} t^{-1}\log\boldsymbol{E}\exp(\theta S^{\delta\lfloor t/\delta\rfloor}) &\leq \lim_{t\to\infty} t^{-1}\log\boldsymbol{E}\exp(\theta Y(t)) \\ &\leq \liminf_{t\to\infty} t^{-1}\log\boldsymbol{E}\exp(\theta S^{\delta\lceil t/\delta\rceil}). \end{aligned} \tag{4.5}$$

Hence,

$$\lim_{n\to\infty} n^{-1}\log\boldsymbol{E}\exp(\theta S_n^\delta) = \delta^{-1}\psi(\theta). \tag{4.6}$$

Since $\delta^{-1}\psi(\theta) = 0$ if and only if $\psi(\theta) = 0$, the proof is complete.

5. Proof of Theorem 5

By Theorem 3,

$$n^{-1}\log\boldsymbol{E}\exp(-\theta S_n^{u*}) \to \psi_u^*(-\theta) = -\psi_A^{-1}(\theta) \text{ as } n \to \infty$$

for each $\theta > 0$. Now we consider the waiting time and workload in the $G/D/1$ queue with the given point process as the arrival process. We let the deterministic service times have mean 1 and the arrival processes have rate ρ where $0 < \rho < 1$. This requires that we scale the original process.

By Theorem 1, in the customer-stationary case (1.1) holds for each ρ, $0 < \rho < 1$, where $\theta_W^*(\rho)$ satisfies Equation (1.30), i.e.

$$\psi_u(-\theta_W^*(\rho)/\rho) = -\theta_W^*(\rho), \qquad 0 < \rho < 1. \tag{5.1}$$

On the other hand, by Theorem 4 and Proposition 7, in the time-stationary case (1.36) holds for each ρ, $0 < \rho < 1$, where the decay rate $\theta_L^*(\rho)$ satisfies

$$\rho\psi_A(\psi_v(\theta_L^*(\rho))) - \theta_L^*(\rho) = \rho\psi_A(\theta_L^*(\rho)) - \theta_L^*(\rho) = 0, \tag{5.2}$$

because the decay rate function of $A(\rho t)$ is $\rho\psi_A$. However, by (1.29), (5.2) is equivalent to

$$-\psi_u^*(-\theta_L^*(\rho)/\rho) = \psi_A^{-1}(\theta_L^*(\rho)/\rho) = \theta_L^*(\rho), \qquad 0 < \rho < 1. \tag{5.3}$$

By Proposition 8, $\theta_W^*(\rho) = \theta_L^*(\rho)$ for $0 < \rho < 1$. Hence, (5.3) becomes

$$\psi_u^*(-\theta_W^*(\rho)/\rho) = -\theta_W^*(\rho), \qquad 0 < \rho < 1. \tag{5.4}$$

Finally, by Proposition 5, (5.1) and (5.4) imply that $\psi_u = \psi_u^*$.

6. Proof of Theorem 6

By Theorem 5, $\psi_u(-\theta) = -\psi_A^{-1}(\theta)$. Paralleling the proof of Theorem 5, we have

$$\psi_u(-\theta_W^*(\rho)/\rho) = -\psi_v(\theta_W^*(\rho)), \qquad 0 < \rho < 1, \tag{6.1}$$

instead of (5.1) and

$$\rho\psi_A(\psi_v(\theta_L^*(\rho)) - \theta_L^*(\rho) = 0, \qquad 0 < \rho < 1, \tag{6.2}$$

instead of (5.2). However, (6.2) is equivalent to

$$\psi_A^{-1}(\theta_L^*(\rho)/\rho) = \psi_v(\theta_L^*(\rho)), \qquad 0 < \rho < 1. \tag{6.3}$$

Since $\psi_A^{-1}(\theta) = -\psi_u(-\theta)$, (6.3) coincides with (6.1), so that we must have $\theta_L^*(\rho) = \theta_W^*(\rho)$, $0 < \rho < 1$.

7. Proof of Proposition 7

Note that, for any $\epsilon > 0$, there is an n_0 such that

$$\begin{aligned} \boldsymbol{E}\exp(\hat{\theta}I(t)) &= \sum_{n=0}^{\infty} \boldsymbol{E}\exp\left(\hat{\theta}\sum_{i=1}^{n} V_i\right)\boldsymbol{P}(A(t) = n) \\ &\le \sum_{n=0}^{\infty} \exp(n(\psi_v(\hat{\theta}) + \epsilon))\boldsymbol{P}(A(t) = n) + \boldsymbol{E}\exp\left(\hat{\theta}\sum_{i=0}^{n_0} V_i\right) \end{aligned}$$

$$\leq \boldsymbol{E}\exp(\psi_v(\hat{\theta})+\epsilon)A(t)+\boldsymbol{E}\exp\left(\hat{\theta}\sum_{i=0}^{n_0} V_i\right)$$

$$\leq \exp(t\psi_A(\psi_v(\hat{\theta})+\epsilon)+\epsilon)+\boldsymbol{E}\exp\left(\hat{\theta}\sum_{i=0}^{n_0} V_i\right)$$

for t suitably large. Hence,

$$\limsup_{t\to\infty} t^{-1}\log\boldsymbol{E}\exp(\hat{\theta}I(t)) \leq \psi_A(\psi_V(\hat{\theta})+\epsilon)+\epsilon.$$

Since ϵ was arbitrary and ψ_A is continuous at $\psi_v(\hat{\theta})$,

$$\limsup_{t\to\infty} t^{-1}\log\boldsymbol{E}\exp(\hat{\theta}I(t)) \leq \psi_A(\psi_V(\hat{\theta})).$$

The reasoning for the other direction is essentially the same.

8. Proof of Proposition 9

We shall work with characterization (ii) in Proposition 1. Note that

$$W=\sum_{i=1}^{(Q^{\mathrm{a}}-1)^+} V_i+V_{\mathrm{e}}^{\mathrm{a}} \geq \sum_{i=1}^{(Q^{\mathrm{a}}-1)^+} V_i \tag{8.1}$$

and

$$L=\sum_{i=1}^{(Q-1)^+} V_i+V_{\mathrm{e}} \geq \sum_{i=1}^{(Q-1)^+} V_i, \tag{8.2}$$

where $V_{\mathrm{e}}^{\mathrm{a}}$ and V_{e} are the equilibrium residual service times of the customer in service (which in general depend upon Q^{a} and Q). Since the argument is essentially the same for W and L, we henceforth consider only W. To have a useful inequality in the opposite direction, we truncate the service times by setting $V_n^c=\min\{V_n, c\}$, $n\geq 1$. Then

$$W^c \leq \sum_{i=1}^{Q^{\mathrm{ac}}} V_i^c + c. \tag{8.3}$$

From (8.1), we obtain

$$\boldsymbol{E}\exp(\theta W) \geq \boldsymbol{E}(\boldsymbol{E}\exp(\theta V_1))^{Q^{\mathrm{a}}-1},$$

so that $\theta_{Q^{\mathrm{a}}}^* \geq \log\boldsymbol{E}\exp(\theta_W^* V_1)$. (As in Proposition 8, we use the fact that $\theta_W^* < \theta_{V_1}^*$.) From (8.3), we obtain

$$\boldsymbol{E}\exp(\theta W^c) \leq \exp(\theta_c)\boldsymbol{E}(\boldsymbol{E}\exp(\theta V_1^c))^{Q^{\mathrm{ac}}},$$

so that

$$\log\boldsymbol{E}\exp(\theta_{W^c}^* V_1^c) \geq \theta_{Q^{\mathrm{ac}}}^*.$$

Then note that V_1^c, W^c, Q^{ac}, Q^c and L^c increase stochastically to their limits V_1, W, Q^a, Q and L as $c \to \infty$, see Theorems 4, 5 and 8 plus the remark on p. 216 of Whitt (1981). Hence

$$\theta^*_{Q^a} \le \liminf_{c\to\infty} \theta^*_{Q^c} \le \lim_{c\to\infty} \log \boldsymbol{E} \exp(\theta^*_{W^c} V_1^c) = \log \boldsymbol{E} \exp(\theta^*_W V_1) < \infty.$$

9. An example

In this section we give an example.

Example 1. To see that the conditions in Theorem 1 are not necessary for (1.1) or (1.2), consider the $G/G/1$ model with

$$\boldsymbol{P}(U_{2n+1} = 1, U_{2n+2} = 1 + Y_n, V_{2n+1} = 1 + Y_n, V_{2n+2} = 0 \text{ for all } n) = 1/2$$

and

$$\boldsymbol{P}(U_{2n+1} = 1 + Y_n, U_{2n+2} = 1, V_{2n+1} = 0,$$
$$V_{2n+2} = 1 + Y_{n+1} \text{ for all } n) = 1/2,$$

where $\{Y_n\}$ is an i.i.d. sequence of exponential random variables with mean 1. Then $\{U_n, V_n\}$ is stationary with $\boldsymbol{E}V_n = 1 < \boldsymbol{E}U_n = 3/2$, so that $\rho = 2/3$. Moreover, it is easy to see that, for $n \ge 1$,

$$\boldsymbol{P}(W_n > x) = \boldsymbol{P}(W > x) = (1/2)e^{-x}, \qquad x > 0,$$

but

$$\boldsymbol{P}(S_{2n+1} = Y_n, S_{2n+2} = -n \text{ for all } n)$$
$$= \boldsymbol{P}(S_1 = -(n + Y_1), S_{2n+2} = Y_{n+1} - (n + Y_1) \text{ for all } n) = 1/2,$$

so that

$$\boldsymbol{E}\exp(\theta S_{2n+1}) = \tfrac{1}{2}\boldsymbol{E}\exp(\theta Y_1) + \tfrac{1}{2}\boldsymbol{E}\exp(-\theta(n + Y_1)),$$
$$\boldsymbol{E}\exp(\theta S_{2n+2}) = \tfrac{1}{2}\exp(-\theta n) + \tfrac{1}{2}\boldsymbol{E}\exp(\theta(Y_{n+1} - n - Y_1)),$$

and

$$n^{-1} \log \boldsymbol{E} \exp(\theta S_n) \to \psi(\theta) = -\theta/2 \; as \; n \to \infty.$$

Hence, (1.1) and (1.2) hold, but $\psi(\theta^*) = 0$ for $\theta^* = 0$.

Note added in proof: Additional results for multiserver queues related to those mentioned on p. 132 appear in Sadowsky and Szpankowski (1995) and Sadowski (1995).

References

Abate, J. and Whitt, W. (1994) A heavy-traffic expansion for asymptotic decay rates of tail probabilities in multi-channel queues. *Operat. Res. Lett.* To appear.

Abate, J., Choudhury, G. L. and Whitt, W. (1994a) Exponential approximations for tail probabilities of queues, I: waiting times. *Operat. Res.* To appear.

Abate, J., Choudhury, G. L. and Whitt, W. (1994b) Exponential approximations for tail probabilities of queues, II: sojourn time and workload. *Operat. Res.* To appear.

Abate, J., Choudhury, G. L. and Whitt, W. (1994c) Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stoch. Models* **10**. To appear.

Abate, J., Choudhury, G. L. and Whitt, W. (1993) Calculation of the $GI/G/1$ waiting-time distribution and its cumulants from Pollaczek's formulas. *AEÜ* **47**, 311–321.

Addie, R. G. and Zuckerman, M. (1993) An approximation for performance evaluation of stationary single server queues. *IEEE Trans. Commun.*, to appear.

Asmussen, S. (1987) *Applied Probability and Queues.* Wiley, New York.

Asmussen, S. (1989) Risk theory in a Markovian environment. *Scand. Actuarial J.* 69–100.

Asmussen, S. (1993) *Fundamentals of Ruin Probability Theory.* Preliminary draft, Aalborg University, Denmark.

Asmussen, S. and Perry, D. (1992) On cycle maxima, first passage problems and extreme value theory for queues. *Stoch. Models* **8**, 421–458.

Borovkov, A. A. (1976) *Stochastic Processes in Queueing Theory.* Springer-Verlag, New York.

Bucklew, J. A. (1990) *Large Deviations Techniques in Decision, Simulation and Estimation.* Wiley, New York.

Chang, C.-S. (1992) Stability, queue length and delay, part II: stochastic queueing networks. *Proc. 31st IEEE Conf. Decision and Control*, Tucson, AZ, 1005–1010.

Chang, C.-S. (1993) Stability, queue length and delay of deterministic and stochastic queueing networks. IBM T. J. Watson Research Center, Yorktown Heights, NY.

Chang, C.-S., Heidelberger, P., Juneja, S. and Shahabuddin, P. (1992) Effective bandwidth and fast simulation of ATM intree networks. IBM T. J. Watson Research Center, Yorktown Heights, NY.

Choudhury, G. L. and Whitt, W. (1994) Heavy-traffic asymptotic expansions for the asymptotic decay rates in the $BMAP/G/1$ queue. *Stoch. Models.* To appear.

Choudhury, G. L., Lucantoni, D. M. and Whitt, W. (1993) Squeezing the most out of ATM. Submitted.

Choudhury, G. L., Lucantoni, D. M. and Whitt, W. (1994) On the effectiveness of effective bandwidths for admission control in ATM networks. *Proc. 14th Internat. Teletraffic Congr.*, Antibes, France. To appear.

Chung, K. L. (1974) *A Course in Probability Theory*, 2nd edn. Academic Press, New York.

Dembo, A. and Zeitouni, O. (1992) *Large Deviations Techniques and Applications.* A. K. Peters, Wellesley, MA.

Ellis, R. (1984) Large deviations for a general class of random vectors. *Ann. Prob.* **12**, 1–12.

Elwalid, A. I. and Mitra, D. (1993) Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking* **1**, 329–343.

Elwalid, A. I. and Mitra, D. (1994) Markovian arrival and service communication systems: spectral expansions, separability and Kronecker-product forms. AT&T Bell Laboratories, Murray Hill, NJ.

Feller, W. (1971) *An Introduction to Probability Theory and its Applications*, Vol. II, 2nd edn. Wiley, New York.

Franken, P., König, D., Arndt, U. and Schmidt, V. (1981) *Queues and Point Processes.* Akademie-Verlag, Berlin.

Gärtner, J. (1977) On large deviations from the invariant measure. *Theory Prob. Appl.* **22**, 24–39.

Gibbens, R. J. and Hunt, P. J. (1991) Effective bandwidths for the multi-type UAS channel. *QUESTA* **9**, 17–28.

Glynn, P. W. and Whitt, W. (1988a) Ordinary CLT and WLLN versions of $L = \lambda W$. *Math. Operat. Res.* **13**, 674–692.

Glynn, P. W. and Whitt, W. (1988b) An LIL version of $L = \lambda W$. *Math. Operat. Res.* **13**, 693–710.

Glynn, P. W. and Whitt, W. (1994) Large deviations behavior of counting processes and their inverses. *QUESTA*. To appear.

Guerin, R., Ahmadi, H. and Naghshineh, M. (1991) Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J. Sel. Areas Commun.* **9**, 968–981.

Iglehart, D. L. and Whitt, W. (1970) Multiple channel queues in heavy-traffic, II: sequences, networks and batches. *Adv. Appl. Prob.* **2**, 355–369.

IGLEHART, D. L. AND WHITT, W. (1971) The equivalence of functional central limit theorems for counting processes and associated partial sums. *Ann. Math. Statist.* **42**, 1372–1378.

KEILSON, J. (1965) *Greens' Function Methods in Probability Theory.* Griffin, London.

KELLY, F. P. (1991) Effective bandwidths at multi-class queues. *QUESTA* **9**, 5–16.

KINGMAN, J. F. C. (1962) On queues in heavy traffic. *J. R. Statist. Soc.* B **24**, 383–392.

NEUTS, M. F. (1986) The caudal characteristic curve of queues. *Adv. Appl. Prob.* **18**, 221–254.

SHWARTZ, A. AND WEISS, A. (1993) *Large Deviations for Performance Analysis: Queues, Communication and Computing.* In preparation.

SMITH, W. L. (1953) On the distribution of queueing times. *Proc. Camb. Phil. Soc.* **49**, 449–461.

SOHRABY, K. (1993) On the theory of general on-off sources with applications in high speed networks. Proc. *INFOCOM '93*, San Francisco. pp. 401–410.

STERN, T. E. AND ELWALID, A. I. (1991) Analysis of a separable Markov-modulated rate model for information-handling systems. *Adv. Appl. Prob.* **23**, 105–139.

TAKÁCS, L. (1962) *Introduction to the Theory of Queues.* Oxford University Press, New York.

TAKÁCS, L. (1963) The limiting distribution of the virtual waiting time and the queue size for a single server queue with recurrent input and general service times. *Sankhyā* A **25**, 91–100.

TAKÁCS, L. (1967) *Combinatorial Methods in the Theory of Stochastic Processes.* Wiley, New York.

TIJMS, H. C. (1986) *Stochastic Modeling and Analysis: A Computational Approach.* Wiley, New York.

VAN OMMEREN, J. C. W. (1988) Exponential expansion for the tail of the waiting-time probability in the single-server queue with batch arrivals. *Adv. Appl. Prob.* **20**, 880–895.

WHITT, W. (1980) Some useful functions for functional limit theorems. *Math. Operat. Res.* **5**, 67–85.

WHITT, W. (1981) Comparing counting processes and queues. *Adv. Appl. Prob.* **13**, 207–220.

WHITT, W. (1994) Tail probabilities with statistical multiplexing and effective bandwidths for multi-class queues. *Telecommunication Systems*, to appear.

References added in proof

SADOWSKY, J. S. (1995) The probability of large queue lengths and waiting times in a heterogeneous multiserver queue, part II: positive recurrence and logarithmic limits. *Adv. Appl. Prob.* **27**(2). To appear.

SADOWSKY, J. S. AND SZPANKOWSKI, W. (1995) The probability of large queue lengths and waiting times in a heterogeneous multiserver queue, part I: tight limits. *Adv. Appl. Prob.* **27**(2). To appear.