Diffusion Approximations for Queues with Server Vacations

Offer Kella; Ward Whitt

Advances in Applied Probability, Vol. 22, No. 3 (Sep., 1990), 706-729.

Stable URL:
http://links.jstor.org/sici?sici=0001-8678%28199009%2922%3A3%3C706%3ADAFQWS%3E2.0.CO%3B2-Y

Advances in Applied Probability is currently published by Applied Probability Trust.

Adv. Appl. Prob. **22**, 706–729 (1990)
Printed in N. Ireland

DIFFUSION APPROXIMATIONS FOR QUEUES WITH SERVER VACATIONS

OFFER KELLA,* *Yale University*
WARD WHITT,** *AT&T Bell Laboratories*

Abstract

This paper studies the standard single-server queue with unlimited waiting space and the first-in first-out service discipline, modified by having the server take random vacations. In the first model, there is a vacation each time the queue becomes empty, as occurs for high-priority customers with a non-preemptive priority service discipline. Approximations for both the transient and steady-state behavior are developed for the case of relatively long vacations by proving a heavy-traffic limit theorem. If the vacation times increase appropriately as the traffic intensity increases, the workload and queue-length processes converge in distribution to Brownian motion with a negative drift, modified to have a random jump up whenever it hits the origin. In the second model, vacations are generated exogenously. In this case, if both the vacation times and the times between vacations increase appropriately as the traffic intensity increases, then the limit process is reflecting Brownian motion, modified by the addition of an exogenous jump process. The steady-state distributions of these two limiting jump-diffusion processes have decomposition properties previously established for vacation queueing models, i.e., in each case the steady-state distribution is the convolution of two distributions, one of which is the exponential steady-state distribution of the reflecting Brownian motion obtained as the heavy-traffic limit without vacations.

SERVICE INTERRUPTIONS; LIMIT THEOREMS; HEAVY TRAFFIC; STOCHASTIC DECOMPOSITION

1. Introduction and summary

Communication, computer and manufacturing systems have recently generated considerable interest in queueing models in which the server occasionally takes random vacations; see Fuhrmann and Cooper (1985), Doshi (1985), (1986), (1990a,b), Federgruen and Green (1986), Lucantoni et al. (1990), and references cited there. Our purpose here is to develop heavy-traffic diffusion approximations for such models. As with previous diffusion approximations for queues (e.g., see Newell (1982) and Harrison (1985)), these diffusion approximations have considerable applied interest because they provide relatively tractable expressions for quantities of interest in quite general models. (For example, the arrivals need not be Poisson.) The approximations can be applied to general

Received 6 February 1989; revision received 5 June 1989.

* Postal address: Department of Operations Research, Yale University, 84 Trumbull St, New Haven, CT 06520, USA.

** Postal address: Room MH 2C-178, AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974, USA.

stable systems, either directly or with refinements such as interpolations with light-traffic limits; see Burman and Smith (1986), Reiman and Simon (1988) and Whitt (1982), (1989). The diffusion approximations here are of special interest in relation to previous diffusion approximations for queues, because the diffusion approximations here involve jumps. Particularly significant is the opportunity the approximations provide to analyze the transient or time-dependent behavior of the processes, e.g., in the spirit of Abate and Whitt (1987).

In this paper, we consider two models involving a single-server queue with unlimited waiting space and the first-in first-out discipline, one in which the server takes a vacation each time the queue becomes empty (as occurs for high priority customers with non-preemptive priorities), and the other where the vacations are generated by an exogenous process, providing an independent random environment for the queue (as occurs in some models of machine breakdowns). It is well known that the standard heavy-traffic limit theorem for the first vacation model (obtained by letting the traffic intensity approach the critical value 1) is the same as if the server takes no vacations. This limiting behavior nevertheless provides important insight; it shows that the degradation in service due to the vacations is relatively negligible under heavy loads compared to the effect of a high traffic intensity. To obtain a more detailed description of the degradation of service due to vacations, *we let the length of the server vacations increase as the traffic intensity ρ increases.* In particular, the length of the vacation should be of order $(1-\rho)^{-1}$, which is the same order as the mean steady-state workload without vacations. The limit is particularly appropriate as an approximation when the vacations are relatively long, but can be applied in any case, e.g., see the approximation for the steady-state mean workload in (2.18).

We show that normalized versions of the workload, queue-length and waiting-time stochastic processes converge in distribution to non-degenerate limits as $\rho \to 1$. For the first model, the limit process is Brownian motion (BM), modified to have a random jump up whenever the process hits the origin. This limit is of course just BM on $[0, \infty)$ with a jump boundary at the origin instead of the usual reflecting boundary in Iglehart and Whitt (1970) or the sticky boundary in Harrison and Lemoine (1981). (See Harrison (1985) for background on BM and its use to model stochastic flow systems.) For the second model, the limit process is reflected Brownian motion (RBM) modified to have random jumps up at intervals determined by an exogenous vacation process. (The exogenous vacation process specifies the jumps and the intervals between the jumps.) There is some technical interest in these limit theorems because the jumps require working with weak convergence in the function space $D(0, \infty)$, excluding the point 0, with Skorohod's (1956) M_1 topology instead of the customary J_1 topology. (See Billingsley (1968), Whitt (1980) and Pollard (1984) for background on weak convergence of probability measures on function spaces.)

For the second model with an exogenous vacation process, *we inflate the times*

between the vacations as well as the vacations themselves. In particular, the times between vacations should grow like $(1-\rho)^{-2}$ while the vacations grow like $(1-\rho)^{-1}$. Unlike the first model, the heavy-traffic behavior of the second model without inflating the vacations and the times between vacations is not the same as the heavy-traffic behavior of the model without vacations, as can be seen from Fischer (1977), Burman and Smith (1986), Burman (1987a, b), and Asmussen (1988). However, without inflating the vacations, the heavy-traffic limit is RBM with different parameters, but without any jumps. The major contribution here is to point out the relevance of the jump-diffusion processes. The limit processes and associated steady-state distributions obtained here are natural candidates for approximations when there are long rare vacations. An example of a setting in which the limits here should be appropriate is a mean service time of 1 and a traffic intensity of $\rho = 0{\cdot}9$ in the system without vacations, plus a mean vacation time of $2(1-\rho)^{-1} = 20$ and a mean time between vacations of $5(1-\rho)^{-2} = 500$. This range of parameter values was not considered in previous work on this model by Federgruen and Green (1986) and Burman (1987b).

The steady-state distributions of the limit processes are interesting, because they exhibit decomposition properties previously established for special cases of these vacation queueing models; see Doshi (1986). (Recent work by Lucantoni et al. (1990) and Doshi (1990a) extends the decomposition results for the vacation queueing models to approximately the same level of generality assumed here; we only assume a joint functional central limit theorem for the arrival and service processes. Nevertheless, the decomposition results for the jump-diffusion processes here are new, because the jump-diffusion process is not directly a queueing process.) For the first model, the approximating steady-state distribution is the convolution of the exponential steady-state heavy-traffic approximation without vacations and the stationary-excess (or equilibrium-residual-life) distribution associated with the vacation-time distribution; see Theorem 2.2. For the second model, the steady-state distribution is again a convolution of the exponential steady-state heavy-traffic approximation without vacations and another distribution. The other distribution has a positive mass at 0 equal to the steady-state probability that the server would be idle (or, equivalently, one minus the traffic intensity) if the queue received work only from the limiting jump process (not the RBM). The conditional distribution given that it is positive is the convolution of the vacation-time stationary-excess distribution and the steady-state distribution of the embedded Markov chain obtained by looking at the entire limit process just prior to the jumps. An explicit expression is obtained when the jumps occur according to a Poisson process; see the Corollary to Theorem 3.3.

For the general case of the second model, we obtain more tractable approximations for both the process and its steady-state distribution by proving a *second heavy-traffic limit theorem* for the limit process obtained from the first heavy-traffic limit. The second limit process is RBM with an exponential stationary distribution.

The jumps in the first limit process produce a larger diffusion coefficient and a higher steady-state mean in the final RBM than in the RBM without vacations.

The rest of this paper is organized as follows. In Section 2 we specify the first model and establish the limiting behavior for it. In Section 3 we do the same for the second model. In Section 4 we give the proofs. Finally, we make a few concluding remarks in Section 5, including discussion about extensions to multiserver queues and queueing networks. Further results are contained in Kella and Whitt (1991).

2. The first model: vacations when the queue becomes empty

The first model is a standard single-server queue with unlimited waiting room and a first-come first-served discipline, in which the server goes on a vacation for a random period each time the queue becomes empty. To be specific, we assume that the server will take another vacation if there are no arrivals during a vacation, but in our heavy-traffic limit the event of no arrivals during a vacation has negligible probability, so that the model variant without successive vacations has the same heavy-traffic limit.

To establish our heavy-traffic limit, we consider a family of systems indexed by the traffic intensity ρ. We specify the stochastic behavior in terms of three stochastic processes, defined independently of ρ: an *arrival counting process* $\{A(t): t \geqq 0\}$, a sequence of *service times* $\{S_n: n \geqq 1\}$, and a sequence of *vacation times* $\{V_n: n \geqq 1\}$. We assume that the vacation-time sequence $\{V_n\}$ is independent of the other two processes. We also assume that the arrival rate and average service time are both 1, i.e.,

$$\lim_{t\to\infty} t^{-1}A(t) = 1 \quad \text{and} \quad \lim_{n\to\infty} n^{-1}\sum_{k=1}^{n} S_k = 1 \text{ w.p. 1.} \tag{2.1}$$

For the ρth system, we use the *scaled arrival process* $A_\rho(t) = A(\rho t)$, which has arrival rate ρ, and leave the service times unchanged, so that the traffic intensity (or offered load, i.e. the arrival rate times the average service time) is ρ. Thus, for the ρth system, the *total input process* is

$$X_\rho(t) = \sum_{k=1}^{A_\rho(t)} S_k, \qquad t \geqq 0. \tag{2.2}$$

For the ρth system, we let the *scaled vacation times be*

$$V_{\rho n} = (1-\rho)^{-1}V_n, \qquad n \geqq 1, \tag{2.3}$$

so that the length of the vacations is growing with ρ (typically at the same rate as the mean steady-state workload without vacations). Of course, it suffices to have $(1-\rho)V_{\rho n} \Rightarrow V_n$ as $\rho \to 1$ for each n, but (2.3) is a simple sufficient condition, which is adequate for developing approximations. For simplicity, we assume that the queue is initially empty for each ρ. This means that a vacation always begins at time $t = 0$.

Note that $\rho < 1$ is the typical stability condition for the ρth system, i.e., the vacations typically do not alter the conditions for stability, but we have not yet made enough assumptions to guarantee that the standard descriptive queueing processes converge to proper limits as $n \to \infty$ or $t \to \infty$. We study the limiting behavior of the queueing processes (transient behavior) as $\rho \uparrow 1$. We propose using the steady-state limiting distributions (obtained by letting $t \to \infty$) of the limiting processes (obtained by letting $\rho \uparrow 1$) to approximate the steady-state distributions for the ρth system for $\rho < 1$; then we must assume that the steady-state distributions for $\rho < 1$ are well defined.

In order to establish our desired heavy-traffic limits, we assume that the arrival process and service times satisfy a joint functional central limit theorem (FCLT), as in Theorem II.1 of Iglehart and Whitt. In particular, let

$$A'_\rho(t) = (1-\rho)[A(t(1-\rho)^{-2}) - t(1-\rho)^{-2}], \qquad t \geqq 0, \tag{2.4}$$

and

$$S'_\rho(t) = (1-\rho) \sum_{k=1}^{[t(1-\rho)^{-2}]} (S_k - 1), \qquad t \geqq 0, \tag{2.5}$$

where $[x]$ is the greatest integer less than or equal to x. Let $\Rightarrow$ denote weak convergence, as in Billingsley (1968), and let $D([0, \infty), J_1)$ be the space of right-continuous real-valued functions on $[0, \infty)$ with left limits, endowed with the usual Skorohod J_1 topology, as in Whitt (1980) and Pollard (1984).

Basic FCLT assumption. We assume that

$$(A'_\rho, S'_\rho) \Rightarrow (B_1, B_2) \quad \text{in} \quad D([0, \infty), J_1) \times D([0, \infty), J_1) \quad \text{as} \quad \rho \to 1,$$

where (B_1, B_2) is two-dimensional Brownian motion (BM) with $(B_1(1), B_2(1))$ having a bivariate normal distribution with mean $(0, 0)$ and covariance matrix $\Sigma = (\sigma_{ij})$.

As discussed in Iglehart and Whitt, there are many conditions under which the basic FCLT assumption holds. The standard sufficient condition is to have the arrival process independent of the service times, the interarrival times associated with $A(t)$ i.i.d. (independent and identically distributed) with finite squared coefficient of variation (variance divided by the square of the mean) c_a^2 and the service times i.i.d. with finite squared coefficient of variation c_s^2. Then $\sigma_{11} = c_a^2$, $\sigma_{22} = c_s^2$ and $\sigma_{12} = \sigma_{21} = 0$. However, neither independence between the processes nor independence within each process is necessary. For example, all three kinds of dependence occur for multiclass queues with independent classes each of which has arrival and service processes satisfying the independence conditions above; see Iglehart and Whitt (1970) and Fendick et al. (1989).

A serious complication for analyzing vacation models with non-Poisson arrival processes and $\rho < 1$ is the residual interarrival time edge effect at the beginning and the end of each vacation. However, in heavy traffic edge effects are asymptotically negligible. To see this, let U_k be the kth interarrival time associated with $A(t)$, so

that $\rho^{-1}U_k$ is the kth scaled interarrival time associated with $A_\rho(t)$. Let

$$U'_\rho(t) = (1-\rho) \sum_{k=1}^{[t(1-\rho)^{-2}]} (\rho^{-1}U_k - \rho^{-1}), \qquad t \geqq 0. \tag{2.6}$$

By the basic FCLT assumption and Theorem 7.3 of Whitt (1980), $U'_\rho \Rightarrow -B_1$ in $D([0, \infty), J_1)$ as $\rho \to 1$. By applying the continuous mapping theorem (Theorem 5.1 of Billingsley) with the maximum jump function $J_c : (D[0, c], R) \to R$, defined by

$$J_c(x) = \sup \{x(t) - x(t-) : 0 \leqq t \leqq c\}, \tag{2.7}$$

which is continuous, we see that

$$(1-\rho) \max \{\rho^{-1}U_k : 1 \leqq k \leqq t(1-\rho)^{-2}\} \Rightarrow 0 \quad \text{as} \quad \rho \to 1. \tag{2.8}$$

However, we shall not directly apply (2.7) and (2.8) in our proofs.

We shall explicitly treat only the continuous-time workload process (the amount of remaining work in service time in the system at time t). The same result (same normalization and same limit process) holds for the queue-length process by a minor variation of the same argument (see Remark 5.2). For the workload process, it is natural to use the argument in Whitt (1971) based on the random sum (2.2), whereas for the queue-length process it is natural to use the argument in Iglehart and Whitt. The same result also holds for the discrete-time embedded processes obtained by looking at these processes just before arrivals (e.g., waiting times), as in Section I.7 of Iglehart and Whitt. By (2.1), the limiting arrival rate is 1, so there is no rescaling in the random time change.

Let $W_\rho(t)$ and $W_{v\rho}(t)$ be the *workload processes* in the ρth system with and without vacations, respectively, and let the associated *normalized processes* be

$$L_\rho(t) = (1-\rho)W_\rho(t(1-\rho)^{-2}) \text{ and } L_{v\rho}(t) = (1-\rho)W_{v\rho}(t(1-\rho)^{-2}), \qquad t \geqq 0. \tag{2.9}$$

For the limits with vacations we use the space $D((0, \infty), M_1)$ instead of $D([0, \infty), J_1)$; i.e., we exclude the point 0 and we change the topology from Skorohod's (1956) J_1 topology used in Billingsley for $D[0, 1]$ to his weaker M_1 topology (convergence J_1 implies convergence M_1). We do not work with the closed interval $[0, \infty)$, because then pointwise convergence must hold at 0, which we will not have. For each ρ, $L_{v\rho}(0) = 0$ but for the limit $L_v(0) = V_1$. However, we have no difficulty if we exclude the origin.

Convergence of a sequence of deterministic functions $\{x_n\}$ to a limit x in $D((0, \infty), T)$, where $T = J_1$ or M_1, is equivalent to convergence of the restrictions in $D([a, b], T)$ for all but countably many pairs (a, b) with $0 < a < b < \infty$ (except at points of discontinuities of x). When the limit function is continuous, convergence in $D([a, b], T)$ is equivalent to uniform convergence. As shown by Kolmogorov (1956), Skorohod (1956) and Pomarede (1976), the J_1 and M_1 topologies both are Polish (metrizable as complete separable metric spaces) and can be characterized by

uniform convergence of parameteric representations of the graphs. For $D([a, b],, M_1)$ we use the complete graph containing all pairs (t, e) in $[a, b] \times R$ such that $x(t-) \leqq e \leqq x(t)$; for $D([a, b], J_1)$ we use the incomplete graph containing all pairs (t, e) in $[a, b] \times R$ such that $e = x(t)$ or $x(t-)$, i.e., the closure of the graph $\{(t, x(t)): a \leqq t \leqq b\}$. For the complete graphs associated with the M_1 topology, a parametric representation is a continuous function $(\tau(t), \chi(t))$ mapping $[a, b]$ onto the complete graph such that $\tau(t)$ is non-decreasing. Convergence $x_n \to x$ as $n \to \infty$ in $D([a, b], M_1)$ holds if there exist parametric representations (τ_n, χ_n) of x_n and (τ, χ) of x such that

$$\lim_{n\to\infty} \max\left\{ \sup_{a\leqq t\leqq b} |\tau_n(t) - \tau(t)|, \sup_{a\leqq t\leqq b} |\chi_n(t) - \chi(t)| \right\} = 0. \tag{2.10}$$

For functions to be close, the J_1 topology requires nearly the same jumps at nearly the same places, whereas the M_1 topology only requires that parametric representations of the complete graphs of the functions be nearly the same. Hence, a function with several closely spaced small jumps up can be close to a function with one large jump up in M_1 but not in J_1. For example, with I_A the indicator function of the set A,

$$I_{[1,1+n^{-1})} + 2I_{[1+n^{-1},2)} \to 2I_{[1,2)} \quad \text{in} \quad D([0, \infty), M_1) \quad \text{as} \quad n \to \infty, \tag{2.11}$$

but not in $D([0, \infty), J_1)$.

In the following theorem we state results for both L_ρ and $L_{v\rho}$ to make comparison easy. The result without vacations follows from Iglehart and Whitt (1970) or Whitt (1971); alternatively, apply Theorems 5.1 and 6.4 of Whitt (1980) plus the basic FCLT assumption. The result with vacations is proved in Section 4.

Recall that the joint distribution of the vacations is arbitrary. To be sure that the jumps up can keep the Brownian motion non-negative, we assume that $\sum_{k=1}^{\infty} V_k = \infty$ w.p. 1.

Theorem 2.1. (a) $L_\rho \Rightarrow R$ in $D([0, \infty), J_1)$ as $\rho \to 1$, where $R \equiv R(t; -1, \sigma^2)$ is reflected Brownian motion (RBM) with drift -1 and diffusion coefficient $\sigma^2 = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$ starting at 0.

(b) $L_{v\rho} \Rightarrow L_v$ in $D((0, \infty), M_1)$ as $\rho \to 1$, where $L_v \equiv L_v(t; B, \{V_n\})$ is Brownian motion $B \equiv B(t; -1, \sigma^2)$ with drift -1 and diffusion coefficient $\sigma^2 = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$, starting at 0, modified to have a jump up each time the process hits 0, with the nth jump being V_n.

Remark 2.1. It is not difficult to show that $L_v \Rightarrow R$ in $D([0, \infty), J_1)$ if $V_n \xrightarrow{p} 0$ for each n. The standard heavy-traffic limit theorem for vacation models in which the vacations do not grow as ρ increases is essentially equivalent to choosing the limiting vacation times V_n very small, which yields the limit R obtained without vacations.

Recall that we have assumed that the queue is initially empty, so that the first vacation $V_{\rho 1}$ begins at $t = 0$. Let $T_{\rho n}$ be the interval between the end of the nth

vacation and the beginning of the $(n+1)$th vacation, and let $C_{\rho n} = V_{\rho n} + T_{\rho n}$ be the nth *busy cycle*, $n \geqq 1$. Let Φ be the standard normal c.d.f. (mean 0 and variance1). Limits for $(T_{\rho 1}, \cdots, T_{\rho n})$ and $(C_{\rho 1}, \cdots, C_{\rho n})$ as $\rho \to 1$ follow immediately from Theorem 2.1 by the continuous mapping theorem (Theorem 5.1 of Billingsley) applied to first-passage times; see Section 7 of Whitt (1980). In fact, these limits are established as part of the proof of Theorem 2.1. See pp. 565, 581 of Abate and Whitt (1987) for more about the limiting inverse Gaussian distribution.

Corollary. As $\rho \to 1$,

$$\{(1-\rho)^2 C_{\rho n}, n \geqq 1\} \Rightarrow \{C_n, n \geqq 1\} \quad \text{in} \quad R^\infty,$$

where C_n, $n \geqq 1$, are mutually independent conditional on $\{V_n\}$, and C_n is the first passage time for $B(t; -1, \sigma^2)$ from V_n to 0. Conditional on V_n, C_n thus has the inverse Gaussian distribution, i.e.,

$$\begin{aligned} P(\sigma^{-2} C_n \leqq t \mid \sigma^{-2} V_n = x) &= P(\inf\{t \geqq 0 \colon B(t; -1, 1) \leqq -x\} \leqq t) \\ &= \Phi\left(\frac{t-x}{\sqrt{t}}\right) + e^{2x} \Phi\left(\frac{-t-x}{\sqrt{t}}\right), \end{aligned} \tag{2.12}$$

so that $E(\sigma^{2k} C_n^k \mid \sigma^{-2} V_n = x) = x$ for $k = 1$, $x + x^2$ for $k = 2$ and $3x + 3x^2 + x^3$ for $k = 3$; and $E(C_n) = E(V_n)$ and $E(C_n^2) = \sigma^{-2} E(V_n) + E(V_n^2)$.

Recall that $R(t) \Rightarrow R(\infty)$ as $t \to \infty$, where $R(\infty)$ has an exponential distribution with mean $\sigma^2/2$. Let $\overset{\mathrm{d}}{=}$ denote equality in distribution.

Theorem 2.2. Suppose that $\{V_n\}$ is i.i.d. with $E(V_1) = m_1 < \infty$. Then $L_{\mathrm{v}}(t) \Rightarrow L_{\mathrm{v}}(\infty)$ as $t \to \infty$, where $L_{\mathrm{v}}(\infty) \overset{\mathrm{d}}{=} R(\infty) + V_{\mathrm{s}}$, $R(\infty)$ and V_{s} are independent and V_{s} has the stationary-excess distribution of V_1, i.e.,

$$P(V_{\mathrm{s}} \leqq x) = m_1^{-1} \int_0^x P(V_1 > y)\, dy. \tag{2.13}$$

Recall that the moments of V_{s} in (2.13) and V_1, are related by

$$E[V_{\mathrm{s}}^k] = E[V_1^{k+1}]/E[V_1](k+1); \tag{2.14}$$

see p. 64 of Cox (1972). Hence, we have the following. Let $m_k = E(V_1^k)$.

Corollary. Under the assumptions of Theorem 2.2,

$$E[L_{\mathrm{v}}(\infty)] = \frac{\sigma^2}{2} + \frac{m_2}{2m_1} \tag{2.15}$$

and

$$\mathrm{Var}\,[L_{\mathrm{v}}(\infty)] = \left(\frac{\sigma^2}{2}\right)^2 + \frac{4m_1 m_3 - 3m_2^2}{12m_1^2}. \tag{2.16}$$

The resulting direct heavy-traffic approximation for $W_{\mathrm{v}\rho}(\infty)$ is of course $(1-\rho)^{-1} L_{\mathrm{v}}(\infty)$ with $V_{\mathrm{s}} = (1-\rho) V_{\rho \mathrm{s}}$, where $V_{\rho \mathrm{s}}$ is the stationary-excess variable as-

sociated with the vacation time $V_{\rho 1}$ in the ρth system, as in (2.13). (Since $V_{\rho 1} = (1-\rho)^{-1}V_1$, $V_{\rho s} = (1-\rho)^{-1}V_s$.) Hence, the *direct heavy-traffic approximation* for $W_{v\rho}(\infty)$ is

$$W_{v\rho}(\infty) \approx \frac{R(\infty)}{1-\rho} + V_{\rho s}. \tag{2.17}$$

As in Whitt (1982), Burman (1987b) and Burman and Smith (1986), we can also consider refinements to the direct heavy-traffic approximation in (2.17). For example, it is natural to replace $(1-\rho)^{-1}R(\infty)$ in (2.17) by a random variable that is 0 with probability $(1-\rho)$ and $(1-\rho)^{-1}R(\infty)$ with probability ρ, so that (2.17) would be exact for the $M/M/1$ model without vacations. Since the jump should correspond to the amount of work to arrive in a vacation time, it is also natural to multiply $V_{\rho s}$ in (2.17) by ρ, and possibly re-introduce some of the variability of the total input process X_ρ. However, we do not study approximation refinements here. We conclude this section by giving our initial refined heavy-traffic approximation for the steady-state mean under the assumptions of Theorem 2.2, namely,

$$E[W_{v\rho}(\infty)] \approx \frac{\rho\sigma^2}{2(1-\rho)} + \rho E[V_{\rho 1}]\frac{(c_v^2+1)}{2}, \tag{2.18}$$

where c_v^2 is the squared coefficient of variation of the vacation time $V_{\rho 1}$. (Recall that the mean service time is 1.) If the model without vacations is $GI/G/1$, then $\sigma^2 = c_a^2 + c_s^2$.

3. The second model: exogenous vacations

Now we assume that the vacations occur exogenously. In addition to the framework of Section 2, this model requires another basic stochastic process, a sequence of non-negative random variables $\{T_n : n \geq 0\}$ with $T_0 = 0$ and $P(T_n > 0) = 1$ for all $n \geq 1$. For the ρth system, we assume that the time between the end of the nth vacation and the beginning of the $(n+1)$th vacation is $T_{\rho n}$, where

$$T_{\rho n} = (1-\rho)^{-2}T_n, \qquad n \geq 1. \tag{3.1}$$

We assume that the vector-valued vacation process $\{(V_n, T_n) : n \geq 1\}$ is independent of $\{A(t) : t \geq 0\}$ and $\{S_n : n \geq 1\}$, which is what we mean by exogenous. As before, we assume that the queue is initially empty and the first vacation begins at time 0. Unlike the first model, the stability criterion for the ρth system of the second model is typically not $\rho < 1$; assuming that all the processes are stationary, the stability condition is typically $\rho < 1 - E(V_{\rho n})/E(T_{\rho n})$. However, by (2.3) and (3.1), $EV_{\rho n}/ET_{\rho n} = (1-\rho)(EV_n/ET_n)$, so that the typical stability condition translates into $\rho < 1$ and $EV_n < ET_n$ here. As in Section 2, we have not made assumptions guaranteeing that steady-state limits exist under these conditions, so we must assume that steady-state distributions exist when we develop approximations for

them. However, approximations will only be developed for the case $\rho < 1$ and $EV_n < ET_n$.

Let $W_{e\rho}$ be the workload process with the exogenous vacations and let $L_{e\rho}$ be the associated normalized workload process, defined just as in (2.9). To describe the limit process, let

$$N(t) = \max\{n \geqq 0 : T_0 + \cdots + T_{n-1} \leqq t\}, \qquad t \geqq 0, \tag{3.2}$$

where $T_0 = 0$. Let the limiting net input process be

$$Y_e(t) = B(t; -1, \sigma^2) + \sum_{i=1}^{N(t)} V_i, \qquad t \geqq 0, \tag{3.3}$$

and then apply a reflecting barrier to get L_e, i.e.,

$$L_e(t) = Y_e(t) + \max\left\{0, -\inf_{0 \leqq s \leqq t} Y_e(s)\right\}, \qquad t \geqq 0; \tag{3.4}$$

see pp. 14, 19 of Harrison (1985). (The reflection is applied after we add the jumps.)

Theorem 3.1. $L_{e\rho} \Rightarrow L_e$ in $D((0, \infty), M_1)$ as $\rho \to 1$, where $L_e \equiv L_e(t; R, \{(V_n, T_n)\})$ is RBM $R(t; -1, \sigma^2)$ modified by having jumps up of size V_n at $T_0 + \cdots + T_{n-1}$, $n \geqq 1$.

The limit process L_e is relatively complicated, as can be seen by observing that if we let the RBM variance σ^2 be 0, then L_e coincides with the workload process in a $G/G/1$ queue with interarrival-time sequence $\{T_n\}$ and service-time sequence $\{V_n\}$. Hence, without extra assumptions on the exogenous vacation process $\{(V_n, T_n)\}$, we cannot expect to obtain very tractable expressions.

To describe the steady-state behavior, we assume that $\{(V_n, T_n)\}$ is i.i.d. Our description of the steady-state distribution of the limit process L_e involves the discrete-time process obtained by looking at the process L_e just prior to the jumps; i.e., let

$$J_e(n) = L_e((T_1 + \cdots + T_n) -), \qquad n \geqq 1. \tag{3.5}$$

For background on the theory of discrete-time real-valued Markov chains, see p. 150 of Asmussen (1987) and Laslett et al. (1978).

Theorem 3.2. If $\{(V_n, T_n) : n \geqq 1\}$ is a sequence of i.i.d. random vectors with $E(V_1) < E(T_1) < \infty$, then $\{J_e(n) : n \geqq 1\}$ is an aperiodic ϕ-irreducible (with ϕ Lebesgue measure) Harris-recurrent Markov chain with an absolutely continuous transition kernel having a strictly positive continuous density; i.e.,

$$P(x, B) \equiv P(J_e(n+1) \in B \mid J_e(n) = x) = \int_B f(x, y)\, dy, \tag{3.6}$$

where $f(x, y)$ is strictly positive and continuous in (x, y).

Corollary. Under the assumptions of Theorem 3.2, it is possible to construct the processes J_e and L_e together with renewal processes such that the processes J_e and L_e become regenerative with successive regeneration cycles being i.i.d. with finite mean.

Theorem 3.3. Under the assumptions of Theorem 3.2, if the distribution of T_1 is non-lattice, then $L_e(t) \Rightarrow L_e(\infty)$ as $t \to \infty$. If, in addition, V_1 is independent of T_1 with $E[V_1^2] < \infty$, then $L_e(\infty) \stackrel{d}{=} R(\infty) + Z$, where $R(\infty)$ and Z are independent, $R(\infty)$ has the limiting exponential distribution of RBM with mean $\sigma^2/2$, $P(Z > 0) = E(V_1)/E(T_1)$, $(Z \mid Z > 0) \stackrel{d}{=} J_e(\infty) + V_s$ with $J_e(\infty)$ and V_s independent, V_s in (2.13) and $J_e(\infty)$ having the stationary distribution of J_e.

Remark 3.1. For the special case in which $\sigma^2 = 0$, the expression for $L_e(\infty)$ in Theorem 3.3 coincides with the known result for the $GI/G/1$ queue; see (3.3) on p. 189 of Asmussen (1987).

We can apply PASTA (see Wolff (1982)) to obtain the following explicit expression when T_1 has an exponential distribution.

Corollary. If, in addition, $P(T_1 \leqq t) = 1 - e^{-\lambda t}$, $t \geqq 0$, then

$$L_e(\infty) \stackrel{d}{=} J_e(\infty) \stackrel{d}{=} R(\infty) + Z',$$

where $R(\infty)$ and Z' are independent, $R(\infty)$ is just as in Theorem 3.3, and

$$E[e^{\alpha Z'}] = \frac{1 - (EV_1/ET_1)}{1 - (EV_1/ET_1)E[e^{-\alpha R(\infty)}]E[e^{-\alpha V_s}]}, \tag{3.7}$$

so that

$$P(Z' \leqq x) = (1 - EV_1/ET_1)\Big(1 + \sum_{n=1}^{\infty} (EV_1/ET_1)^n (F_1^{n^*} * F_2^{n^*})\Big)(x) \tag{3.8}$$

where $F_i^{n^*}$ is the n-fold convolution of the exponential distribution of $R(\infty)$ for $i = 1$ and of $P(V_s \leqq x)$ for $i = 2$, and

$$E[Z'] = \frac{(EV_1/ET_1)(ER(\infty) + EV_s)}{(1 - (EV_1)/(ET_1))}. \tag{3.9}$$

Remark 3.2. When T_1 has an exponential distribution, the net input process associated with $L_{e\rho}$ is a Lévy process without negative jumps, so that (3.7)–(3.9) can also be calculated directly from Harrison (1977). See Kella and Whitt (1991).

Remark 3.3. When $\sigma^2 = 0$, $R(\infty) = 0$ and (3.7) reduces to the classical Pollaczek–Khintchine formula for the $M/G/1$ queue; see p. 206 of Asmussen (1987).

If T does not have an exponential distribution, then we do not yet have a tractable expression for the steady-state variable $L_e(\infty)$. Hence we prove a *second heavy-traffic limit* for the limit process L_e. (A similar limit could be established for the first model, but there is little motivation.) For this purpose, we construct a family of

processes indexed by η with $\eta < 1$. Let $N(t)$ be the counting process associated with $\{T_n\}$ in (3.2). Paralleling (2.4) and (2.5), let

$$N'_\eta(t) = (1-\eta)[N(t(1-\eta)^{-2}) - \lambda t(1-\eta)^{-2}], \qquad t \geqq 0, \tag{3.10}$$

and

$$V'_\eta(t) = (1-\eta) \sum_{k=1}^{[t(1-\eta)^{-2}]} (V_k - \nu), \qquad t \geqq 0. \tag{3.11}$$

For each η, let the process N depend on η through simple time scaling as for A_ρ in Section 2, so that $\lambda_3 = \eta/\nu$, i.e., η represents the growth rate of the input of work associated with vacations, as indicated by the translation terms. We then make another FCLT assumption.

Second FCLT assumption. We assume that

$$(N'_\eta, V'_\eta) \Rightarrow (\tilde{B}_1, \tilde{B}_2) \quad \text{in} \quad D([0,\infty), J_1) \times D([0,\infty), J_1) \quad \text{as} \quad \eta \to 1,$$

where $(\tilde{B}_1, \tilde{B}_2)$ is two-dimensional BM with $(\tilde{B}_1(1), \tilde{B}_2(1))$ having a bivariate normal distribution with mean vector $(0,0)$ and covariance matrix $\tilde{\Sigma} = (\tilde{\sigma}_{ij})$.

Finally, let $L'_{e\eta}$ be the normalized process associated with the jump-diffusion process L_e, i.e.,

$$L'_{e\eta}(t) = (1-\eta)L_e(t(1-\eta)^{-2}), \qquad t \geqq 0. \tag{3.12}$$

Theorem 3.4. Under the second FCLT assumption with $\eta = \lambda/\nu$,

$$L'_{e\eta} \Rightarrow R' \quad \text{in} \quad D([0,\infty), J_1) \quad \text{as} \quad \eta \to 1,$$

where $R' \equiv R'(t; -1, \sigma^2 + \nu\tilde{\sigma}^2)$ is RBM with $\sigma^2 = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$ as in Theorem 3.1, ν is the centering constant in (3.11) and $\tilde{\sigma}^2 = \tilde{\sigma}_{11} + \tilde{\sigma}_{22} - 2\tilde{\sigma}_{12}$.

Note that $R'(\infty)$ has an exponential distribution with mean $(\sigma^2 + \nu\tilde{\sigma}^2)/2$. The resulting direct two-stage heavy-traffic approximation for the steady-state workload $W_{e\rho\eta}(\infty)$ is

$$W_{e\rho\eta}(\infty) \approx (1-\rho)^{-1} L_{e\eta}(\infty) \approx (1-\rho)^{-1}(1-\eta)^{-1} R'(\infty), \tag{3.13}$$

where

$$\eta = \frac{EV_1}{ET_1} = \frac{EV_{\rho 1}}{ET_{\rho 1}(1-\rho)}, \tag{3.14}$$

so that $(1-\rho)^{-1}(1-\eta)^{-1} = 1 - \rho - (EV_{\rho 1}/ET_{\rho 1})$ and

$$E[W_{e\rho\eta}(\infty)] \approx \frac{\sigma^2 + \nu\tilde{\sigma}^2}{2(1-\rho-(EV_{\rho 1}/ET_{\rho 1}))}. \tag{3.15}$$

However, for practical applications it appears that (3.13) and (3.14) are much in need of refinement. From Theorem 3.3 and its Corollary, it is evident that we should regard $(1-\eta)^{-1}R'(\infty)$ only as an approximation for the difficult component

$(Z \mid Z>0)$ in Theorem 3.3. Theorems 3.3 and 3.4 thus suggest that the distributional form for an approximation of $W_{e\rho}(\infty)$ be the convolution of two distributions, each of which is the mixture of an exponential and a point mass at 0. As a practical refined approximation, we propose

$$W_{e\rho}(\infty) \approx Z_1 + Z_2, \tag{3.16}$$

where Z_1 and Z_2 are independent, $P(Z_1 = 0) = 1 - \rho$, $P(Z_2 = 0) = (1 - \rho - \delta)/(1 - \rho)$ where $\delta = EV_{\rho 1}/(EV_{\rho 1} + ET_{\rho 1})$ is the long-run proportion of time the server is on vacation, $(Z_i \mid Z_i > 0)$ is exponential for $i = 1, 2$,

$$E(Z_1 \mid Z_1 > 0) = \frac{\sigma^2}{2(1-\rho)} \tag{3.17}$$

and

$$E(Z_2 \mid Z_2 > 0) = \frac{\rho(\sigma^2 + \nu\tilde{\sigma}^2)}{2(1-\rho-\delta)}. \tag{3.18}$$

The random variable Z_1 is a familiar approximation for $W_{e\rho}(\infty)$ without any vacations, as suggested by Theorem 3.3. We know that the exact probability that the server is idle and not on vacation is $1 - \rho - \delta$; this is the basis for our approximation of $P(Z_2 = 0)$. Finally, the approximation for $(Z_2 \mid Z_2 > 0)$ is based on Theorem 3.4.

The resulting approximation for the mean is

$$E[W_{e\rho}(\infty)] \approx \frac{\rho\sigma^2}{2(1-\rho)} + \frac{\rho\delta(\sigma^2 + \nu\tilde{\sigma}^2)}{2(1-\rho)(1-\rho-\delta)}. \tag{3.19}$$

When the model is $GI/G/1$ with $\{V_n\}$ and $\{T_n\}$ independent sequences of i.i.d. random variables with squared coefficients of variation c_v^2 and c_t^2, (3.19) becomes

$$E[W_{e\rho}(\infty)] \approx \frac{\rho(c_a^2 + c_s^2)}{2(1-\rho)} + \frac{\rho\delta(c_a^2 + c_s^2 + \nu(c_v^2 + c_t^2))}{2(1-\rho)(1-\rho-\delta)} \tag{3.20}$$

where $\nu = EV_{\rho 1}$.

Alternatively, the conditional mean $E(Z_2 \mid Z_2 > 0)$ can be chosen so that the approximate overall mean $E[W_{e\rho}(\infty)]$ matches a separately determined approximation, such as the interpolation approximation for the mean number in system in (3.3) of Burman (1987b), which has proven to be an effective approximation in comparisons with simulations. (The mean number in system can be obtained from the mean delay using Little's formula.) For the $M/G/1$ case, Burman's interpolation approximation is (in our notation)

$$\begin{aligned} E[W_{e\rho}(\infty)] \approx{}& \frac{\rho(1-\rho-\delta)\delta\nu(c_s^2+1)}{2(1-\delta)} \\ &+ \frac{\rho}{1-\delta}\left[1 + \frac{\rho(1+c_s^2)}{2(1-\rho-\delta)}\right]\left[1 + \rho\delta\nu\left(\frac{c_y^2 + c_t^2}{1+c_s^2}\right)\right]. \end{aligned} \tag{3.21}$$

Our limit involves $\delta \to 0$, $\nu \to \infty$ and $\delta\nu \to (EV_1)^2/(EV_1 + ET_1)$ as $\rho \to 1$.

4. Proofs

In this section we prove the previous theorems.

Proof of Theorem 2.1(*b*). We establish weak convergence by considering the process over successive busy cycles. Let $L_{v\rho n}$ be the normalized process restricted to the nth busy cycle, i.e.,

$$L_{v\rho n}(t) = L_{v\rho}(t)1_{[U_{\rho,n-1},U_{\rho n})}(t(1-\rho)^{-2}), \qquad t \geqq 0, \tag{4.1}$$

where $U_{\rho n} = C_{\rho 1} + \cdots + C_{\rho n}$ with $U_{\rho 0} = 0$ and, as before, $1_A(t)$ is the indicator function of the set A, so that

$$L_{v\rho}(t) = \sum_{n=1}^{\infty} L_{v\rho n}(t), \qquad t \geqq 0. \tag{4.2}$$

By considering deterministic functions (sample paths), we see that $L_{v\rho} \Rightarrow L_v \equiv \sum_{n=1}^{\infty} L_{vn}$ as $\rho \to 1$ in $D(0, \infty)$ with any of the Skorohod topologies if

$$(L_{v\rho 1}, \cdots, L_{v\rho n}, (1-\rho)^2 C_{\rho 1}, \cdots, (1-\rho)^2 C_{\rho n}) \Rightarrow (L_{v1}, \cdots, L_{vn}, C_1, \cdots, C_n) \quad \text{as} \quad \rho \to 1 \tag{4.3}$$

in $D(0, \infty)^n \times R^n$ with the same topology on D, where $P(C_n > 0) = 1$ for each n, and $C_1 + \cdots + C_n \xrightarrow{p} \infty$ as $n \to \infty$. Of course, here L_{vn} is BM $B(t; -1, \sigma^2)$ starting at V_n at time $C_1 + \cdots + C_{n-1}$ absorbing at 0 and $C_1 + \cdots + C_n$ is the time that it is absorbed. Since we have assumed that $\sum_{n=1}^{\infty} V_n = \infty$ w.p. 1, $(C_1 + \cdots + C_n) \xrightarrow{p} \infty$ as $n \to \infty$.

We proceed by mathematical induction on the cycle index, establishing the corollary to Theorem 2.1 along the way. For cycle n, we will establish four results. First, the normalized length of the vacation $V_{\rho n}$ converges in probability to 0; second, the normalized work to arrive during $V_{\rho n}$ converges in probability to V_n; third, the normalized workload process after the vacation (without further vacations) converges weakly to Brownian motion starting at V_n; fourth, the time until the workload next becomes empty converges weakly to the first-passage time of BM to 0 starting in V_n.

First, $(1-\rho)^2 V_{\rho n} \xrightarrow{p} 0$ as $\rho \to 1$ for each n by (2.3). Second, the amount of work to arrive in the nth vacation is $X_\rho(U_{\rho,n-1} + V_{\rho n}) - X_\rho(U_{\rho,n-1})$. However, by the basic FCLT assumption and Theorem 5.1 of Whitt (1980) or Section 17 of Billingsley (1968), $X'_\rho \Rightarrow B'$ in $D([0, \infty), J_1)$ as $\rho \to 1$, where

$$X'_\rho(t) = (1-\rho)[X_\rho(t(1-\rho)^{-2}) - \rho t(1-\rho)^{-2}], \qquad t \geqq 0, \tag{4.4}$$

and $B' \equiv B'(t; 0, \sigma^2)$ is BM with drift coefficient 0 and diffusion coefficient $\sigma^2 = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$. By the induction hypothesis,

$$(1-\rho)^2 U_{\rho,n-1} \Rightarrow U_{n-1} \equiv C_1 + \cdots + C_{n-1} \quad \text{in} \quad R \quad \text{as} \quad \rho \to 1. \tag{4.5}$$

Hence,

$$\begin{aligned}(1-\rho)[X_\rho(U_{\rho,n-1} + V_{\rho n}) - X_\rho(U_{\rho,n-1})] &= X'_\rho((1-\rho)^2 U_{\rho,n-1} \\ &+ (1-\rho)^2 V_{\rho n}) - X'_\rho((1-\rho)^2 U_{\rho n}) + \rho(1-\rho)V_{\rho n} \xrightarrow{p} V_n \quad \text{as} \quad \rho \to 1.\end{aligned} \tag{4.6}$$

In this last step we used the convergence of X'_ρ, (2.3), (4.5) and the continuous mapping theorem: if $y_n \to y$ in $[0, \infty)$ and $x_n \to x$ in $D([0, \infty), J_1)$ as $n \to \infty$, where x has continuous paths, then $x_n(y_n) \to x(y)$ in R. Thus, both X'_ρ terms in (4.6) converge weakly to $B'(U_{n-1})$, so that their difference converges in probability to 0.

Third, let Y_ρ be the net input process, defined by $Y_\rho(t) = X_\rho(t) - t$, $t \geqq 0$, and let Y'_ρ be the associated normalized process

$$Y'_\rho(t) = (1-\rho)Y_\rho(t(1-\rho)^{-2}) = X'_\rho(t) - t, \qquad t \geqq 0, \tag{4.7}$$

so that $Y'_\rho \Rightarrow B' - e$ as $\rho \to 1$, where $e(t) = t$, $t \geqq 0$. Then the normalized process on $[U_{\rho,n-1} + V_{\rho n}, \infty)$, before being absorbed when it hits 0, is

$$\begin{aligned}(1-\rho)V_{\rho n} + Y'_\rho((1-\rho)^2(U_{\rho,n-1} + V_{\rho n}) + \cdot) - Y'_\rho(U_{\rho,n-1} + V_{\rho n}) \\ \Rightarrow V_n + (B' - e)(U_{n-1} + \cdot) - (B' - e)(U_{n-1}) \stackrel{\mathrm{d}}{=} V_n + B' - e.\end{aligned} \tag{4.8}$$

Fourth, by the continuous mapping theorem, the first-passage times to 0 converge weakly too, so that $(1-\rho)^2 C_{\rho n} \Rightarrow C_n$ and (4.5) holds for n as well as $n-1$.

Finally, we put the pieces together to conclude that (4.3) holds for each n if we use the M_1 topology on $D((0, \infty))$. We obtain convergence of the successive n-dimensional processes by considering joint limits with previously established limits, as in Whitt (1971). Since the ρth process is non-decreasing in each vacation, we have weak convergence to the limit process with the jump in the M_1 topology for $n \geqq 2$. For $n = 1$, we have excluded the limiting jump by considering the space $D(0, \infty)$.

Proof of Theorem 2.2. Since the BM has negative drift and $E(V_1) < \infty$, the jump epochs are regeneration points for the process L_v. By the Corollary to Theorem 2.1, the mean interval between regeneration points is $E(V_1) < \infty$. Moreover, since the regeneration interval has a positive density, $L_v(t) \Rightarrow L_v(\infty)$ as $t \to \infty$. It thus remains to determine the distribution of $L_v(\infty)$. For this purpose, note that the limit process L_v starting at $x \geqq 0$ can be represented as

$$L_v(t) = x - t + \sigma B(t) + \sum_{i=1}^{N(t)} V_i, \qquad t \geqq 0, \tag{4.9}$$

where $\sigma^2 = c_a^2 + c_s^2$, $B(t) \equiv B(t; 0, 1)$ is standard Brownian motion,

$$\begin{aligned}N(t) &= \sup\{n \geqq 0 : \tau_n \leqq t\}, \qquad t \geqq 0, \\ \tau_n &= \inf\left\{t \geqq 0 : x - t + \sigma B(t) + \sum_{i=0}^{n-1} V_i = 0\right\}, \qquad n \geqq 1,\end{aligned} \tag{4.10}$$

$V_0 = 0$ and $\tau_0 = 0$; i.e., L_v can be represented as the stochastic integral

$$L_v(t) = L_v(0) + \int_0^t \sigma \, dB(s) + Y(t), \qquad t \geqq 0, \tag{4.11}$$

where

$$Y(t) = -t + \sum_{i=0}^{N(t)} V_i, \qquad t \geqq 0; \tag{4.12}$$

($L_v(t)$ is adapted to the standard filtration $\mathcal{F}_t$ generated by $\{B(s): 0 \leqq s \leqq t\}$ and $\{V_n: n \geqq 1\}$; see Chapter 4 of Harrison (1985) and Chung and Williams (1983). The fact that we include all of $\{V_n\}$ in $\mathcal{F}_t$ for each t shows that independence in $\{V_n\}$ is not essential.) Hence, we can apply the generalized Ito formula on p. 71 of Harrison (1985) and p. 301 of Meyer (1976) with the function $f(u) = e^{-\alpha u}$, $\alpha > 0$, to obtain

$$\begin{aligned} \exp(-\alpha L_v(t)) = \exp(-\alpha x) - \alpha\sigma \int_0^t \exp(-\alpha L_v(s))\, dB(s) \\ + \left(\alpha + \frac{\alpha^2\sigma^2}{2}\right) \int_0^t \exp(-\alpha L_v(s))\, ds - \sum_{i=0}^{N(t)} (1 - \exp(-\alpha V_i)). \end{aligned} \tag{4.13}$$

Since $L_v(t) \geqq 0$, $\exp(-\alpha L_v(t))$ is a bounded process, so that the stochastic integral in (4.13) is a continuous L_2 martingale with mean 0; see p. 62 of Harrison (1985) or p. 40 of Chung and Williams. Thus, taking expected values in (4.13) and using Tonelli's theorem, we obtain

$$\begin{aligned} E[\exp(-\alpha L_v(t))] = e^{-\alpha x} - E \sum_{i=0}^{N(t)} (1 - \exp(-\alpha V_i)) \\ + \left(\alpha + \frac{\alpha^2\sigma^2}{2}\right) \int_0^t E[\exp(-\alpha L_v(s))]\, ds. \end{aligned} \tag{4.14}$$

Since $L_v(t) \Rightarrow L_v(\infty)$ as $t \to \infty$, $E[\exp(-\alpha L_v(t))]$ converges to a proper limit. From the regenerative structure and the Corollary to Theorem 2.1, $t^{-1}N(t) \to EV_1$ w.p. 1 and

$$\begin{aligned} t^{-1} \sum_{i=1}^{N(t)} (1 - \exp(-\alpha V_i)) \\ = \frac{N(t)}{t} \frac{1}{N(t)} \sum_{i=1}^{N(t)} (1 - \exp(-\alpha V_i)) \to \frac{(1 - E[\exp(-\alpha V_1)])}{EV_1} \text{ w.p. 1.} \end{aligned}$$

Since there exists a constant M such that

$$E\left[t^{-1} \sum_{i=1}^{N(t)} (1 - \exp(-\alpha V_i))\right]^2 \leqq t^{-2} E[N(t)^2] \leqq M < \infty \tag{4.15}$$

for $t \geqq 1$ (e.g., see the argument in the proof of (12) on p. 136 of Chung), the process on the left in (4.15) is uniformly integrable (p. 95 of Chung), so that the

means converge too. Hence, dividing by t in (4.14) and letting $t \to \infty$, we obtain

$$\begin{aligned} \lim_{t\to\infty} E\exp(-\alpha L_v(t)) &= \lim_{t\to\infty} t^{-1}\int_0^t E[\exp(-\alpha L_v(s))]\,ds \\ &= \left(\frac{2/\sigma^2}{2/\sigma^2+\alpha}\right)\left(\frac{1-E\exp(-\alpha V_1)}{\alpha E V_1}\right), \end{aligned} \tag{4.16}$$

which is the product of the Laplace transforms of the appropriate two distributions.

Proof of Theorem 3.1. The argument is almost identical to the proof of Theorem 2.1(b), but slightly easier because we can analyze the vacations separately before considering the queue. With the scaling in (2.3) and (3.1), $(1-\rho)^2 T_{\rho n} \Rightarrow T_n$ and $(1-\rho)^2 V_{\rho n} \Rightarrow 0$, so that the analog of the Corollary to Theorem 2.1 holds and the limit process has jumps at the times $T_1 + \cdots + T_n$ for $n \geqq 1$. We then consider successive cycles, just as in the proof of Theorem 2.1.

Proof of Theorem 3.2. Observe that J_e is a Markov chain with transition kernel $P(x, B)$ in (3.6), where

$$f(x, y) = E[g(x + J_1, T_1, y)] \tag{4.17}$$

with $g(x, t, y)$ being the density of RBM at time t starting in x; p. 49 of Harrison (1985) or (1.1) of Abate and Whitt (1987). Since g is a strictly positive and continuous function of (x, t, y), $f(x, y)$ is a strictly positive and continuous function of (x, y). Thus, $P(x, B) > 0$ for all Borel sets B with $\phi(B) > 0$ where ϕ is Lebesgue measure; i.e., J_e is aperiodic and ϕ-irreducible; p. 457 of Laslett et al. (1978). Moreover, J_e is weakly continuous ($Ph(x)$ is continuous for each bounded continuous real-valued h); p. 459 of Laslett et al. (1978) and p. 224 of Billingsley (1968). To establish recurrence, we apply the mean drift criterion; see Theorem 2.2 of Laslett et al. (1978). Since the transition kernel of RBM with density g is stochastically monotone (p. 564 of Abate and Whitt (1987)), so is $P(x, B)$. Since $EV_1 < ET_1 < \infty$ and the RBM has drift -1, for any $\epsilon > 0$ there exists K such that

$$E(J_e(n+1) \mid J_e(n) = x) \leqq \begin{cases} x - \epsilon & \text{for } x > K \\ K - \epsilon & \text{for } x \leqq K, \end{cases} \tag{4.18}$$

so that

$$\sup\{E[\inf\{n \geqq 1 : X_n \leqq K\} \mid X_0 = x] : 0 \leqq x \leqq K\} < \infty \tag{4.19}$$

and any set of positive Lebesgue measure is recurrent. Finally, J_e is Harris recurrent by Example 3.1 on p. 151 of Asmussen (1987); $r = 1$ there.

Proof of the Corollary to Theorem 3.2. The regenerative structure follows immediately from Harris recurrence; p. 151 of Asmussen. In general, the regeneration cycles are only 1-dependent, but here they are independent because we have $r = 1$.

Proof of Theorem 3.3. The regenerative structure provided by the Corollary to Theorem 3.2 and the assumption that T_1 has a non-lattice distribution imply that $L_e(t) \Rightarrow L_e(\infty)$ as $t \to \infty$; apply Proposition 3.2 on p. 187 of Asmussen to deduce that the regenerative cycle distribution in continuous time is also non-lattice and then apply Theorem 1.2 on p. 126. To calculate the distribution of $L_e(\infty)$, note that the process L_e starting in 0 can be represented as

$$L_e(t) = \sigma B(t) + U(t) - t + Y(t), \qquad t \geqq 0, \tag{4.20}$$

where B is (0, 1) BM,

$$U(t) = \sum_{i=1}^{N(t)} V_i, \qquad t \geqq 0, \tag{4.21}$$

$$N(t) = \sup\{n \geqq 0 : T_0 + \cdots + T_{n-1} \leqq t\}, \qquad t \geqq 0, \tag{4.22}$$

$T_0 = 0$ and

$$Y(t) = \max\{0, -\inf\{\sigma B(s) + U(s) - s : 0 \leqq s \leqq t\}\}, \qquad t \geqq 0, \tag{4.23}$$

as on pp. 14–20 of Harrison (1985). Just as in the proof of Theorem 2.2, we apply the generalized Ito formula on p. 71 of Harrison (1985) with the function $f(u) = e^{-\alpha u}$, here obtaining

$$\begin{aligned} \exp(-\alpha L_e(t)) = \exp(-\alpha V_1) - \alpha\sigma \int_0^t \exp(-\alpha L_e(s))\, dB \\ + \left(\alpha + \frac{\alpha^2\sigma^2}{2}\right) \int_0^t \exp(-\alpha L_e(s))\, ds \\ - \alpha Y(t) + \sum_{i=1}^{N(t)} \exp(-\alpha J_e(i))(\exp(-\alpha V_{i+1}) - 1). \end{aligned} \tag{4.24}$$

As in the proof of Theorem 2.2, the stochastic integral with respect to BM in (4.24) is an L_2 martingale. Hence, we can take expected values, divide by t and let $t \to \infty$ to obtain

$$E[\exp(-\alpha L_e(\infty))] = \lim_{t\to\infty} t^{-1} \int_0^t E[\exp(-\alpha L_e(s))]\, ds$$

$$= \left(\alpha + \frac{\alpha^2\sigma^2}{2}\right)^{-1} \lim_{t\to\infty} \left(\alpha t^{-1} EY(t) + (1 - E[\exp(-\alpha V_1)]) E\left[t^{-1} \sum_{i=1}^{N(t)} \exp(-\alpha J_e(i))\right]\right). \tag{4.25}$$

As in the proof of Theorem 2.2, by the regenerative structure,

$$t^{-1} \sum_{i=1}^{N(t)} \exp(-\alpha J_e(i)) \to \frac{E[\exp(-\alpha J_e(\infty))]}{E[T_1]} \quad \text{w.p. 1} \quad \text{as} \quad t \to \infty \tag{4.26}$$

and the expected values converge too due to uniform integrability, because $E[(N(t)/t)^2]$ is uniformly bounded, as in (4.15).

By the strong law of large numbers, $t^{-1}B(t) \to 0$ and $t^{-1}U(t) \to EV_1/ET_1$ w.p. 1 as $t \to \infty$. Hence,

$$t^{-1}Y(t) \to 1 - \frac{EV_1}{ET_1} \quad \text{w.p. 1} \quad \text{as} \quad t \to \infty. \tag{4.27}$$

One way to prove (4.27) is to note that the strong law is equivalent to a functional strong law (Theorem 4 of Glynn and Whitt (1988)) and then apply Theorem 6.2(ii) of Whitt (1980). It remains to show that $t^{-1}Y(t)$ is uniformly integrable. By Minkowski's inequality (p. 47 of Chung),

$$E[(t^{-1}Y(t))^2]^{\frac{1}{2}} \leqq \left(t^{-2}\sigma^2 E\left[\left(\sup_{0 \leqq s \leqq t} B(s)\right)^2\right]\right)^{\frac{1}{2}} + (t^{-1})^{\frac{1}{2}} + (t^{-2}E[U(t)^2])^{\frac{1}{2}}. \tag{4.28}$$

Since $\sup_{0 \leqq s \leqq t} B(s) \overset{\mathrm{d}}{=} |B(t)|$ (p. 8 of Harrison (1985)), $E[(\sup_{0 \leqq s \leqq t} B(s))^2] = t$. By conditioning on $N(t)$ in (4.21) and then unconditioning,

$$t^{-2}E[U(t)^2] = t^{-2}E[V_1^2]E[N(t)] + t^{-2}(EV_1)^2(E[N(t)] + E[N(t)^2]). \tag{4.29}$$

Since $E[V_1^2] < \infty$ and $t^{-2}E[N(t)^2]$ is uniformly bounded for $t \geqq 1$, $t^{-2}E[U(t)^2]$ is uniformly bounded.

Finally, combining (4.25)–(4.27), we obtain

$$\begin{aligned} E[\exp(-\alpha L_e(\infty))] = \left(\frac{2/\sigma^2}{2/\sigma^2 + \alpha}\right)\Bigg[\left(1 - \frac{EV_1}{ET_1}\right)1 \\ + \left(\frac{EV_1}{ET_1}\right)\left(\frac{1 - E[\exp(-\alpha V_1)]}{\alpha EV_1}\right)E[\exp(-\alpha J_e(\infty))]\Bigg] \end{aligned} \tag{4.30}$$

as claimed.

Proof of Theorem 3.4. As in (2.2) and (4.4), let

$$\tilde{X}'_\eta(t) = (1 - \eta)[X_\eta(t(1 - \eta)^{-2}) - \eta t(1 - \eta)^{-2}], \qquad t \geqq 0, \tag{4.31}$$

where

$$X_\eta(t) = \sum_{k=1}^{N_\eta(t)} V_k, \qquad t \geqq 0. \tag{4.32}$$

By the second FCLT assumption, $\tilde{X}'_\eta \Rightarrow \tilde{B} \equiv B(t; 0, \nu\tilde{\sigma}^2)$ is BM with drift coefficient 0 and diffusion coefficient $\nu\tilde{\sigma}^2 = \nu(\tilde{\sigma}_{11} + \tilde{\sigma}_{22} - 2\tilde{\sigma}_{12})$, where ν is the centering term in (3.11). Since

$$\{(1 - \eta)B(t(1 - \eta)^2; 0, \sigma^2) : t \geqq 0\} \overset{\mathrm{d}}{=} \{B(t; 0, \sigma^2) : t \geqq 0\}, \tag{4.33}$$

the net input process associated with $L'_{e\eta}$ in (3.12) converges to $B \equiv B(t; -1, \sigma^2 + \nu\tilde{\sigma}^2)$. Finally, the desired convergence of $L'_{e\eta}$ itself follows by the continuous mapping theorem with the barrier mapping, i.e., Theorem 6.4 of Whitt (1980).

5. Concluding remarks

5.1. *Limits for the steady-state distributions.* The *stochastic decomposition property* for queues with vacations says that a steady-state characteristic such as the waiting time is distributed as the sum of two independent random variables, one of which is the corresponding steady-state characteristic without vacations; see p. 37 of Doshi (1986). Theorems 2.2 and 3.3 are new results of this form for the steady-state distribution of jump-diffusion processes. Given that such a decomposition property has been established for a queueing model, with a tractable expression for the second component, it is usually relatively easy to establish heavy-traffic limits for the steady-state distribution directly. For example, the distribution of $L_v(\infty)$ in Theorem 2.2 can be obtained by taking limits with the $M/G/1/V_M$ model steady-state quantities in (4.4) and (4.17) of Doshi (1986). We can often treat the steady-state limits without vacations by applying the heavy-traffic results of Kingman (1962).

5.2. *The queue-length process.* As indicated earlier, the queue-length process can be treated much like the workload process. For example, consider the proof of Theorem 3.1b. Instead of the amount of work to arrive in the nth vacation, we use the number of arrivals to arrive in the nth vacation, $A_\rho(U_{\rho,n-1} + V_{\rho n}) - A_\rho(U_{\rho,n-1})$. For each interval between vacations, instead of the net input process Y_ρ, we use the arrival counting process minus the potential service counting process, where the potential service counting process is the counting process associated with the service times $\{S_n\}$ evaluated at the cumulative busy time of the server. The cumulative busy time up to time $U_{\rho,n-1} + V_{\rho n} + t$ before the $(n+1)$th vacation is

$$U_{\rho,n-1} - (V_{\rho 1} + \cdots + V_{\rho,n-1}) + t.$$

Hence, the scaled net input process representing the arrivals minus the services after the nth vacation converges to BM with drift coefficient -1 and diffusion coefficient $\sigma^2 = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$ as $\rho \to 1$.

5.3. *Conditions on* $\{V_n\}$ *in Theorem* 2.2. In Theorem 2.2 it is not necessary for the vacation times to be mutually independent. For example, a minor modification of the current proof applies if the vacation sequence $\{V_n\}$ is regenerative with $E(V_1) < \infty$. In the new proof of Theorem 2.2, regeneration points for the process L_v are the jump epochs corresponding to regeneration points in the vacation sequence. If $\{V_n\}$ is only stationary, then $\{L_v(t): t \geqq 0\}$ has a stationary version with the same stationary marginal distribution as if $\{V_n\}$ is i.i.d. To see this, note that the time between the nth and $(n+1)$th jump in L_v conditioned on V_n has an inverse Gaussian distribution with mean and variance V_n. For Theorem 2.2 it is also not necessary for the vacation time sequence $\{V_n\}$ to be independent of the Brownian motion. Our proof based on Ito's formula requires that a jump V_n occurring at time t not depend on the process L_v after t (i.e., the random sum of all the jumps up to time t is adapted to the σ-field $\mathcal{F}_t$ in the filtration). Under regularity conditions guaranteeing that convergence holds, which can be seen from our proof, the

distribution of $L_v(\infty)$ is just as stated above when the two assumed kinds of independence are relaxed.

5.4. *Conditions on* $\{(V_n, T_n)\}$ *in Theorem* 3.3. In the proof of Theorem 3.3, the condition $E[V_1^2] < \infty$ and the condition that V_n be independent of T_n for each n are used only to establish uniform integrability of $t^{-1}U(t)$ in (4.21) and (4.29) and thus $t^{-1}Y(t)$ in (4.23) and (4.28). Hence, it is apparent that these conditions can be relaxed. Moreover, as in Remark 5.3, the generalized Ito formula used in the proof of Theorem 3.3 does not require that the vacation vectors (V_n, T_n) be mutually independent or independent of the Brownian motion. It suffices for $U(t) = \sum_{i=1}^{N(t)} V_i$ to be adapted to the σ-field $\mathcal{F}_t$ in the filtration for the process L_e. However, the limiting behavior of the expectation of the final term in (4.24) becomes more complicated.

5.5. *Multiserver queues.* The results for the first model extend easily to multiserver queues if all servers go on vacation together the instant one server becomes idle, but the heavy-traffic behavior is much more complicated if the servers go on vacation separately. Then the number of servers that go on vacation, which determines the size of the jump, seems to depend on the more detailed structure of the process.

On the other hand, for the second model it is relatively straightforward to extend the results to multiple servers with a general exogenous vacation process and the same scaling. Assuming s homogeneous servers (which is not necessary), there will be a jump up of $(k/s)V_n$ if k servers simultaneously go on vacation for a period of $V_{\rho n}$ satisfying (2.3). *An interesting feature is that the generalization of Theorem* 3.1 *requires working with Skorohod's* (1956) M_2 *topology* instead of the M_1 topology used with one server. The M_2 topology can be characterized in terms of the Hausdorff metric applied to the complete graphs; see Pomarede (1976). We need the M_2 topology because the converging process does not have only many small jumps up where it approaches the limiting jump up, due to the fact that some servers are still working while others are on vacation. We know of no previous application of the M_2 topology.

5.6. *Open queueing networks.* Finally, we observe that the heavy-traffic limits can be extended to queueing networks. A key step is to observe that we easily obtain heavy-traffic limits for the departure processes in our models from our results. In particular, with either model, the departure counting process in the ρth system can be represented as $D_\rho(t) = A_\rho(t) - Q_\rho(t)$ where $Q_\rho(t)$ is the number of customers in the ρth system. Hence, a FCLT for the normalized version of D_ρ follows from the joint FCLT for (A_ρ, Q_ρ), which holds by a minor modification of our arguments, see Remark 5.2 and Iglehart and Whitt (1970); it is easy to obtain the joint limit as well as the limits for the components separately. To treat the subtraction, apply the M_1 analog of Theorem 4.1 in Whitt (1980).

The departure process D_ρ is interesting because it is a point process with large gaps. The resulting limit process is complicated, just as without vacations, but it has continuous paths except for jumps of size V_n down at the times the limit process for the queue-length process has jumps of size V_n up.

The FCLT limit theorem for the departure process implies a corresponding FCLT for the queue-length process at a subsequent single-server queue, just as in II, Theorem 1 and Section 4 of Iglehart and Whitt (1970). The jumps down in the limit process for the departure process from the first queue cause a simultaneous jump down in the limit process for the queue length at the second queue. If the size of the jump exceeds the value of the queue-length limit process before the jump, the excess will appear as a gap in the limit process for the departure process from the second queue, and so forth. If the departures are routed randomly to two or more different queues with a Markovian transition matrix P_{ij}, then a jump down of V_n in the limit process for the departure process from queue i results in a jump down of $V_n P_{ij}$ in the limit process for the flow from queue i to the queue j (essentially by the law of large numbers). If there is feedback to the queue with the initial jump up, then some of the inital jump up may be 'cancelled' by simultaneous jumps down caused by the gap in the departure process. In general, each vacation causes an instantaneous vector-valued jump transition in the vector-valued limit processes associated with the queue lengths or the workloads at all the queues. Overall, we can obtain weak convergence of the normalized version of the vector-valued workload and queue-length processes to vector-valued jump-diffusion processes in cases already treated without vacations. Paralleling the proof of Theorem 2.1, we use previous arguments in Reiman (1984) and Chen and Mandelbaum (1988) to treat the vector-valued processes in the intervals between vacations and modifications of the proof of Theorem 2.1 here, plus the rough arguments above, to treat the vacations (the jumps). (More details are intended for a subsequent paper.)

We thus can define generalizations of the Brownian networks in Harrison and Reiman (1981), Reiman (1984), Harrison and Williams (1987), Harrison (1988) and Chen and Mandelbaum (1988), in which there are instantaneous jump transitions, in the manner sketched above. These extensions evidently make complicated limit processes even more complicated, but the behavior of the jumps in the network model seems to provide useful insight. In the heavy-traffic time scale, a relatively long (but not too long) server vacation has an instantaneous effect on the entire network.

References

ABATE, J. AND WHITT, W. (1987) The transient behavior of regulated Brownian motion, I and II. *Adv. Appl. Prob.* **19,** 560–631.

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

ASMUSSEN, S. (1988) The heavy traffic limit of a class of Markovian queueing models. *Operat. Res. Letters* **6,** 301–306.

BILLINGSLEY, P. (1968) *Convergence of Probability Measures*. Wiley, New York.

BURMAN, D. Y. (1987a) Diffusion approximations for queueing systems an analytic approach. AT&T Bell Laboratories, Murray Hill, N.J.

BURMAN, D. Y. (1987b) Approximations for a service system with interruptions. AT&T Bell Laboratories, Murray Hill, N.J.

BURMAN, D. Y. AND SMITH, D. R. (1986) An asymptotic analysis of a queueing system with Markov-modulated arrivals. *Operat. Res.* **34,** 105–119.

CHEN, H. AND MANDELBAUM, A. (1988) Stochastic discrete flow networks: diffusion approximations and bottlenecks. Graduate School of Business, Stanford University.

CHUNG, K. L. (1974) *A course in Probability Theory*. 2nd edition. Academic Press, New York.

CHUNG, K. L. AND WILLIAMS, R. J. (1983) *Introduction to Stochastic Integration*. Birkhäuser, Boston.

COX, D. R. (1972) *Renewal Theory*. Methuen, London.

DOSHI, B. T. (1985) A note on stochastic decomposition in a $GI/G/1$ queue with vacations or set-up times. *J. Appl. Prob.* **22,** 419–428.

DOSHI, B. T. (1986) Queueing systems with vacations—a survey. *Queueing Systems* **1,** 29–66.

DOSHI, B. T. (1990a) Generalizations of the stochastic decomposition results for single server queues with vacations. *Stochastic Models*. To appear.

DOSHI, B. T. (1990b) Single server queues with vacations. In *Stochastic Analysis of Computer and Communications Systems*, ed. H. Takagi, North-Holland, Amsterdam, 217–265.

FEDERGRUEN, A. AND GREEN, L. (1986) Queueing systems with service interruptions. *Operat. Res.* **34,** 752–768.

FENDICK, K. W., SAKSENA, V. R. AND WHITT, W. (1989) Dependence in packet queues. *IEEE Trans. Commun.*, **37,** 1173–1183.

FISCHER, M. J. (1977) An approximation to queueing systems with interruptions. *Management Sci.* **24,** 338–344.

FUHRMANN, S. AND COOPER, R. B. (1985) Stochastic decompositions in an $M/G/1$ queue with generalized vacations. *Operat. Res.* **33,** 1117–1129.

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

HARRISON, J. M. (1977) The supremum distribution of a Lévy process with no negative jumps. *Adv. Appl. Prob.* **9,** 417–422.

HARRISON, J. M. (1985) *Brownian Motion and Stochastic Flow Systems*. Wiley, New York.

HARRISON, J. M. (1988) Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Applications*. ed. W. Fleming and P. L. Lions, Springer-Verlag, New York, 147–186.

HARRISON, J. M. AND LEMOINE, A. J. (1981) Sticky Brownian motion as the limit of a storage process. *J. Appl. Prob.* **18,** 216–226.

HARRISON, J. M. AND REIMAN, M. I. (1981) Reflected Brownian motion on the orthant. *Ann. Prob.* **9,** 302–308.

HARRISON, J. M. AND WILLIAMS, R. (1987) Brownian models of open queueing networks with homogeneous customer populations. *Stochastics* **22,** 77–115.

IGLEHART, D. L. AND WHITT, W. (1970) Multiple channel queues in heavy traffic, I and II. *Adv. Appl. Prob.* **2,** 150–177 and 355–369.

KELLA, O. AND WHITT, W. (1991) Queues with server vacations and Lévy processes with secondary jump inputs. *Ann. Appl. Prob.* **1**. To appear.

KINGMAN, J. F. C. (1962) On queues in heavy traffic. *J. R. Statist. Soc.* **B24,** 383–392.

KOLMOGOROV, A. N. (1956) On Skorohod convergence. *Theory Prob. Appl.* **1,** 215–222.

LASLETT, G. M., POLLARD, D. B. AND TWEEDIE, R. L. (1978) Techniques for establishing ergodic and recurrence properties of continuous-valued Markov chains. *Naval Res. Log. Quart.* **25,** 455–472.

LUCANTONI, D., MEIER-HELLSTERN, K. AND NEUTS, M. F. (1990) A single server queue with server vacations and a class of non-renewal arrival processes. *Adv. Appl. Prob.* **22,** 676–705.

MEYER, P.-A. (1976) *Un Cours sur les Intégrales Stochastiques*. Lecture Notes in Mathematics **511,** Springer-Verlag, New York.

NEWELL, G. F. (1982) *Applications of Queueing Theory*, 2nd edn. Chapman and Hall, London.

POLLARD, D. (1984) *Convergence of Stochastic Processes*. Springer-Verlag, New York.

POMAREDE, J. L. (1976) A Unified Approach Via Graphs to Skorohod Topologies on the Function Space *D*. Ph.D. Dissertation, Yale University.

REIMAN, M. I. (1984) Open queueing networks in heavy traffic. *Math. Operat. Res.* **9,** 441–458.

REIMAN, M. I. AND SIMON, B. (1988) An interpolation approximation for queueing systems with Poisson input. *Operat. Res.* **36,** 454–469.

SKOROHOD, A. V. (1956) Limit theorems for stochastic processes. *Theory Prob. Appl.* **1,** 261–290.

TAKAGI, H. (1986) *Analysis of Polling Systems*. MIT Press, Cambridge, MA.

WHITT, W. (1971) Weak convergence theorems for priority queues: preemptive-resume discipline. *J. Appl. Prob.* **8,** 74–94.

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

WHITT, W. (1982) Refining diffusion approximations for queues. *Operat. Res. Letters* **1,** 165–169.

WHITT, W. (1989) An interpolation approximation for the mean workload in the $GI/G/1$ queue. *Operat. Res.* **37,** 936–952.

WOLFF, R. W. (1982) Poisson arrivals see time averages. *Operat. Res.* **30,** 223–231.