The Annals of Applied Probability
1995, Vol. 5, No. 3, 681–719

POLLING SYSTEMS WITH ZERO SWITCHOVER TIMES: A HEAVY-TRAFFIC AVERAGING PRINCIPLE

By E. G. Coffman, Jr., A. A. Puhalskii[1] and M. I. Reiman

AT & T Bell Laboratories; Institute for Problems in Information Transmission, Moscow; and AT & T Bell Laboratories

In polling systems, $M \geq 2$ queues are visited by a single server in cyclic order. These systems model such diverse applications as token-ring communication networks and cyclic production systems. We study polling systems with exhaustive service and zero switchover (walk) times. Under standard heavy-traffic assumptions and scalings, the total unfinished work converges to a one-dimensional reflected Brownian motion, whereas the workloads of individual queues change at a rate that becomes infinite in the limit. Although it is impossible to obtain a multidimensional limit process in the usual sense, we obtain an "averaging principle" for the individual workloads. To illustrate the use of this principle, we calculate a heavy-traffic estimate of waiting times.

1. Introduction.

A polling system consists of $M \geq 2$ queues visited by a server in cyclic order. In the traditional system studied here, the server remains at a queue serving customers in first in–first out (FIFO) order until none remains, the case of exhaustive service. When a queue is empty on the server's arrival or when it becomes empty after the server finishes the last waiting customer, the server moves instantaneously to the next queue in sequence. That is, we assume zero switchover (walk) times, thus confining ourselves to applications (e.g., certain of those arising in computer/communication settings) in which this is a useful approximation.

In the stochastic model studied here, independent arrival processes are assumed for the M queues. Each arrival process consists of a sequence of i.i.d. interarrival times drawn from a given general distribution that may vary from one queue to the next. Service times comprise a sequence of i.i.d. random variables independent of interarrival times and with a given general distribution, the same for all queues.

The analysis of polling systems and its many variants has a large and growing literature. For example, see Takagi (1986) and Boxma and Takagi (1992). Polling problems have attracted wide interest not only because of their practical importance, but also because they couple an elegantly simple structure with challenging analysis. The chief obstacle to explicit results is

Received October 1993; revised December 1994.

[1] The research of this author was supported in part under contract with AT & T Bell Laboratories, and in part under a grant from the International Science Foundation (Grant R8200).

AMS 1991 *subject classifications.* 60K25, 60F17, 90B22.

Key words and phrases. Polling systems, cyclic servers, diffusion approximations, heavy-traffic limits.

the interdependence of queueing processes that holds even under simplifying distributional (e.g., exponential) assumptions. A classical Markov-chain approach must adopt a state that carries jointly the states of the M queues. An attempt at explicit formulas for queue-length or waiting-time distributions eventually founders, culminating typically in a system of equations that must be solved numerically.

In these circumstances, one naturally resorts to asymptotic estimates. The touchstone for the success of such techniques lies in the existence of limit laws which show that the estimates are asymptotically exact. Here, we study diffusion approximations in which the asymptotic regime is that of heavy traffic. Reflected Brownian motions (RBM's) approximate the total number in system and the total unfinished work, under the usual heavy-traffic scalings. The theory that can be called upon to support such approximations is well developed. However, for the polling system we consider, in the time scale of the RBM limits the *individual* queue-length and unfinished work processes change at an infinite rate. As a result, the problem of formulating and proving useful limit theorems for the joint distributions seems to be much more difficult.

To illustrate the limit processes analyzed in later sections, consider the symmetric, two-queue ($M = 2$) system and let (V_1, V_2) denote the limiting unfinished work in the two queues under the heavy-traffic normalization. Figure 1 represents the motion of the limit process (V_1, V_2) by a component along the constant-work lines $V = V_1 + V_2$, and an orthogonal component along the diagonal where V varies. While V varies as RBM along the diagonal, (V_1, V_2) moves back and forth along the cross diagonal at an infinite rate, the direction being determined by which of the two queues is being served.

We estimate normalized waiting times as follows. Informally, given $V = V_1 + V_2$ a random arrival finds the process (V_1, V_2) at a point uniformly distributed over the constant-work line V and moving in either direction with equal probability. Thus in the heavy traffic normalization, the state seen by a randomly chosen arrival is taken to be $(UV, (1 - U)V)$, where U is a uniform random variable on $[0, 1]$ independent of V. With probability $1/2$ the arrival's queue is being served, in which case the waiting time is UV in distribution, and with probability $1/2$ the other queue is being served, in which case the waiting time is $UV + (1 - U)V/(1 - \rho/2)$ in distribution, where ρ is the overall traffic intensity. Setting $\rho = 1$ for the heavy-traffic approximation, the normalized waiting time is thus uniform on $[0, V]$ with probability $1/2$ and uniform on $[V, 2V]$ with probability $1/2$, which is distributionally equivalent to being uniform on $[0, 2V]$. Thus we can write

$$Z = 2UV \tag{1.1}$$

for the normalized waiting time.

In Section 2, a general averaging principle is formalized for the case $M = 2$, arbitrary arrival rates and a general service-time distribution, the same for each queue. Appropriately specialized, this principle underlies the

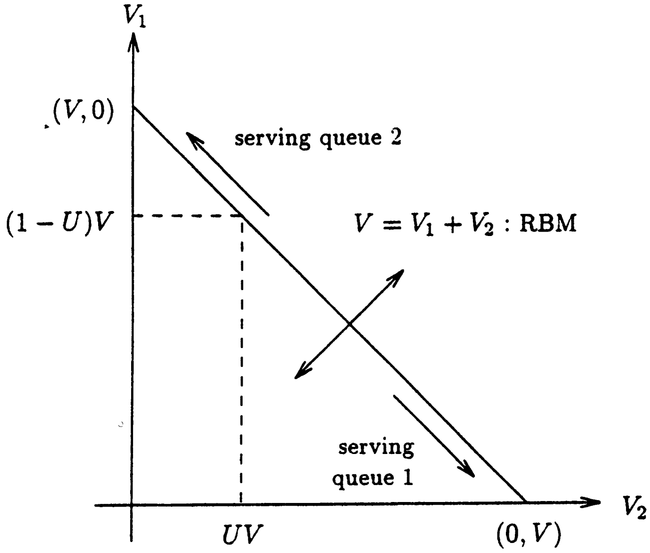

FIG. 1. *The limit process.*

discussion above. Several preliminary results also appear in Section 2. Section 3 prepares for the proof of the averaging principle by analyzing a single-server threshold queue; Section 4 completes the proof. Extensions to the case $M > 2$ and different service-time distributions at each queue are discussed in detail in Section 5, but no proofs are given. Calculations of waiting times are illustrated in Section 6.

2. The heavy-traffic limit, $M = 2$. Consider a sequence of polling systems, each consisting of two queues, with the following parameters for the nth system:

$\{\xi^n_{l,i},\ i \geq 1\}$, $l = 1, 2$, are the (independent) sequences of i.i.d. interarrival times at the lth queue;
$\lambda^n_l = (E\xi_l^n)^{-1}$ and $(\sigma_l^n)^2 = \operatorname{Var}\xi_l^n$, $l = 1, 2$, are the rate and variance parameters of generic interarrival-time random variables, ξ_l^n;
$\{\eta_i^n,\ i \geq 1\}$ is the sequence of i.i.d. service times, assumed to be independent of the arrival processes;
$\mu^n = (E\eta^n)^{-1}$ and $(\sigma_s^n)^2 = \operatorname{Var}\eta^n$ are the rate and variance parameters of a generic service time, η^n.

We assume that the following heavy-traffic conditions hold:

$$\lim_{n\to\infty} \lambda^n_l = \lambda_l > 0, \qquad l = 1, 2, \qquad \lim_{n\to\infty} \mu^n = \mu = \lambda_1 + \lambda_2, \tag{2.1}$$

$$\lim_{n\to\infty} \sqrt{n}\,(\lambda^n_1 + \lambda^n_2 - \mu^n) = c, \tag{2.2}$$

for some finite constant c, and

$$\begin{aligned}\lim_{n\to\infty} \sigma_l^{\,n} &= \sigma_l, \qquad l = 1,2, \qquad \lim_{n\to\infty} \sigma_s^{\,n} = \sigma_s, \\ \sigma^2 &= \lambda_1^3 \sigma_1^2 + \lambda_2^3 \sigma_2^2 + \mu^3 \sigma_s^{\,2} > 0.\end{aligned} \tag{2.3}$$

We also assume that the Lindeberg condition holds:

$$\begin{aligned}&\lim_{n\to\infty} E(\xi_l^n)^2 1\left(\xi_l^n > \varepsilon\sqrt{n}\right) = 0, \qquad l = 1,2, \\ &\lim_{n\to\infty} E(\eta^n)^2 1\left(\eta^n > \varepsilon\sqrt{n}\right) = 0,\end{aligned} \tag{2.4}$$

for all $\varepsilon > 0$. For later use we define $\rho_i = \lambda_i/\mu$, $i = 1,2$, as the limiting traffic intensity in queue i.

Let $Q_l^n(t)$, $t \geq 0$, $l = 1,2$, be the lth queue length and let $Q^n(t) = Q_1^n(t) + Q_2^n(t)$ be the total queue length at time t. Define the normalized processes $X^n = (X^n(t),\ t \geq 0)$ with $X^n(t) = (1/\sqrt{n})Q^n(nt)$. If $X^n(0) \to_P 0$ $(n \to \infty)$, then conditions (2.1)–(2.4) imply that, in the Skorohod space $D[0,\infty)$, X^n converges in distribution to reflected Brownian motion with drift c and diffusion coefficient σ [see Iglehart and Whitt (1970)]. The limit process is denoted by $X = (X(t),\ t \geq 0) = \mathrm{RBM}(c, \sigma^2)$. The central result of this paper is the following *averaging principle* for the normalized, individual queue length, $X_l^n(t) = (1/\sqrt{n})Q_l^n(nt)$, $l = 1,2$.

THEOREM 2.1. *If* $X^n(0) \to_P 0$ *and if conditions* (2.1)–(2.4) *hold, then for any continuous function* f *on* R_+ *and any* $T > 0$,

$$\int_0^T f(X_l^n(t))\, dt \to_d \int_0^T \left(\int_0^1 f(uX(t))\, du \right) dt, \qquad l = 1,2 \text{ as } n \to \infty.$$

REMARK. The integrals above are well defined since $X_l^n(t)$ and $X(t)$ are P-a.s. bounded on $[0, T]$.

Section 3 and the remainder of this section prepare the ground for the proof of Theorem 2.1, which appears in Section 4. Section 3 proves an averaging principle for a special single-server queue; this result plays a key role in the proof of Theorem 2.1. The five lemmas concluding this section are either well known or easily proved. We give them here for ease of future reference. In what follows the notation $\to_d$ will apply both to sequences of random variables and to sequences of stochastic processes; the usage will be clear in context.

LEMMA 2.1. *Let* $X^n = (X^n(t),\ t \geq 0)$ *be a sequence of right-continuous processes with left limits and suppose that* X^n *has paths unbounded above. Assume that, for* $b > 0$, $X^n(t) \to_P bt$ $(n \to \infty)$ *uniformly on finite intervals.*

(i) *Denoting first passage times by* $F^n(t) = \inf(s > 0\colon X^n(s) > t)$, *we have* $F^n(t) \to_P t/b$ $(n \to \infty)$ *uniformly on finite intervals.*

(ii) *Let* $(t^n,\ n \geq 1)$ *be a sequence of times with* $t^n \to t_0$ $(n \to \infty)$ *and define* $F^n = \inf(s > 0\colon X^n(s) > t^n)$. *Then* $F^n \to_P t_0/b$ $(n \to \infty)$.

If the X^n *are increasing, the local uniform convergence in probability of* $X^n(t)$ *to* t *can be replaced by convergence at every* $t \geq 0$.

Result (i) can be found in Iglehart and Whitt (1970) and Whitt (1980). The version dealing with increasing X^n appears in Krichagina, Liptser and Puhalskii (1988). Result (ii) is an obvious consequence of (i).

In the next lemma, bear in mind that if E denotes a metric space, then convergence in distribution in E^∞ is finite-dimensional convergence.

Lemma 2.2. *Let* $X^n = (X_1^n, X_2^n, \dots)$, $n = 1, 2, \dots,$ *be a sequence of random elements of* $D[0,\infty)^\infty$, *where* $X_k^n = (X_k^n(t),\ t \geq 0)$, $k = 1, 2, \dots,$ *are real-valued increasing right-continuous processes. Let* $\tau^n = (\tau_1^n, \tau_2^n, \dots)$, $n \geq 1$, *be a sequence of* R_+^∞*-valued random elements defined on the same probability space. If* $\tau^n \to_d \tau$ *and* $X_k^n \to_P X_k$, *where* $\tau = (\tau_1, \tau_2, \dots)$ *and where* $X_k = (X_k(t),\ t \geq 0)$, $k = 1, 2, \dots,$ *are deterministic and continuous, then* $X^n(\tau^n) \to_d X(\tau)$, *where* $X(\tau) = (X_1(\tau_1), X_2(\tau_2), \dots)$.

Proof. By Theorem 4.4 in Billingsley (1968) we know that $(X^n, \tau^n) \to_d (X, \tau)$ as $n \to \infty$, with convergence being in $D[0,\infty)^\infty \times R_+^\infty$. By the Skorohod embedding we may assume that

$$(X^n, \tau^n) \to (X, \tau) \qquad P\text{-a.s.} \tag{2.5}$$

Then for $\varepsilon > 0$, $\delta > 0$, $T > 0$, $k = 1, 2, \dots,$

$$\begin{aligned} &P(|X_k^n(\tau_k^n) - X_k(\tau_k)| > \varepsilon) \\ &\quad \leq P(|\tau_k^n - \tau_k| > \delta) + P(\tau_k > T) \\ &\qquad + P\left(\sup_{t \leq T + \delta} |X_k^n(t) - X_k(t)| > \frac{\varepsilon}{2}\right) \\ &\qquad + 1\left(\sup_{s, t \leq T + \delta,\, |t - s| \leq \delta} |X_k(t) - X_k(s)| > \frac{\varepsilon}{2}\right). \end{aligned} \tag{2.6}$$

The X_k^n are increasing and X_k is continuous, so by (2.5) the first and third terms on the right of (2.6) tend to 0 as $n \to \infty$. The fourth term is also 0 for δ small enough, by the continuity of X_k. Finally, $P(\tau_k > T) \to 0$ $(T \to \infty)$ takes care of the second term, so by (2.6), $X_k^n(\tau_k^n) \to_P X_k(\tau_k)$ $(n \to \infty)$. By the definition of the topology on R^∞, we have $X^n(\tau^n) \to_P X(\tau)$. The lemma then holds since convergence in probability implies convergence in distribution [Theorem 4.3 of Billingsley (1968)]. □

Lemma 2.3. *For* $\varepsilon > 0$, *let real-valued random variables* $\alpha^n(\varepsilon)$, $\beta^n(\varepsilon)$ *and* γ^n, $n = 1, 2, \dots,$ *satisfy* $\alpha^n(\varepsilon) \leq \gamma^n \leq \beta^n(\varepsilon)$; $\alpha^n(\varepsilon) \to_d \alpha(\varepsilon)$, $\beta^n(\varepsilon) \to_d \beta(\varepsilon)$ $(n \to \infty)$; *and* $\alpha(\varepsilon) \to_d \gamma$, $\beta(\varepsilon) \to_d \gamma$ $(\varepsilon \to 0)$. *Then* $\gamma^n \to_d \gamma$ $(n \to \infty)$.

Proof. Let x be a continuity point of the distribution of γ. Then

$$\begin{aligned}\limsup_{n\to\infty} P(\gamma^n \le x) &\le \limsup_{n\to\infty} P(\alpha^n(\varepsilon) \le x)\\ &\le P(\alpha(\varepsilon) \le x) \to P(\gamma \le x), \qquad \varepsilon \to 0,\end{aligned}$$

so $\limsup_{n\to\infty} P(\gamma^n \le x) \le P(\gamma \le x)$. On the other hand,

$$\begin{aligned}\liminf_{n\to\infty} P(\gamma^n < x) &\ge \liminf_{n\to\infty} P(\beta^n(\varepsilon) < x)\\ &\ge P(\beta(\varepsilon) < x) \to P(\gamma < x), \qquad \varepsilon \to 0.\end{aligned}$$

The lemma follows. □

Lemma 2.4. *Let $\{x_i^n,\ n \ge 1,\ i \ge 1\}$ be a triangular array of random variables. If, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(|x_i^n| > \varepsilon) = 0$$

and

$$\lim_{k\to\infty} \limsup_{n\to\infty} \sum_{i=1}^{n} P(|x_i^n| > k) = 0,$$

then

$$\frac{1}{n} \sum_{i=1}^{n} x_i^n \to_P 0.$$

In particular, if $\{x_i^n,\ i \ge 1\}$ are identically distributed, the above conditions are reduced to

$$x_1^n \to_P 0, \qquad n \to \infty,$$

and

$$\lim_{k\to\infty} \limsup_{n\to\infty} nP(|x_1^n| > k) = 0.$$

Proof. The result follows either from general theorems on the law of large numbers for sums of random variables [e.g., see Petrov (1975)] or directly, since if $\delta > 0$, then for arbitrary $k > \delta$,

$$\begin{aligned}&P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i^n\right| > \delta\right)\\ &\quad\le P\left(\frac{1}{n}\sum_{i=1}^{n} |x_i^n| 1\left(\frac{\delta}{2} \le |x_i^n| \le k\right) > \frac{\delta}{2}\right) + P\left(\sum_{i=1}^{n} 1(|x_i^n| > k) \ge 1\right)\\ &\quad\le P\left(\frac{1}{n}\sum_{i=1}^{n} 1\left(|x_i^n| \ge \frac{\delta}{2}\right) > \frac{\delta}{2k}\right) + \sum_{i=1}^{n} P(|x_i^n| > k)\\ &\quad\le \frac{2k}{\delta}\frac{1}{n}\sum_{i=1}^{n} P\left(|x_i^n| > \frac{\delta}{2}\right) + \sum_{i=1}^{n} P(|x_i^n| > k),\end{aligned}$$

and this last term goes to 0 as $n \to \infty$ and $k \to \infty$ by the assumptions of the lemma. □

The next lemma deals with the continuity of "penetration" times [cf. Jacod and Shiryaev (1987), Section VI.2.11].

LEMMA 2.5. *Let G_1 and G_2 be open subsets of R with nonintersecting closures, and let G_1^ε and G_2^ε denote their open ε-neighborhoods, $\varepsilon \geq 0$, with $G_1^0 = G_1$ and $G_2^0 = G_2$. For $x = (x(t),\ t \geq 0) \in D[0,\infty)$, define $\zeta_0(x, G_2^\varepsilon) = 0$ and, for $k \geq 1$, $T > 0$ ($\inf \varnothing = \infty$),*

$$\tau_k(x, G_1^\varepsilon) = \inf(t > \zeta_{k-1}(x, G_2^\varepsilon) \colon x(t) \in G_1^\varepsilon), \qquad k \geq 1,$$

$$\zeta_k(x, G_2^\varepsilon) = \inf(t > \tau_k(x, G_1^\varepsilon) \colon x(t) \in G_2^\varepsilon), \qquad k \geq 1.$$

If $\hat{x} = (\hat{x}(t),\ t \geq 0) \in D[0,\infty)$ is continuous and, for $k \geq 1$, $T > 0$,

$$\lim_{\varepsilon \downarrow 0} \tau_k(\hat{x}, G_1^\varepsilon) \wedge T = \tau_k(\hat{x}, G_1) \wedge T,$$

$$\lim_{\varepsilon \downarrow 0} \zeta_k(\hat{x}, G_2^\varepsilon) \wedge T = \zeta_k(\hat{x}, G_2) \wedge T,$$

then, as maps from $D[0,\infty)$ to R_+, $x \to \tau_k(x, G_1) \wedge T$ and $x \to \zeta_k(x, G_2) \wedge T$, $k \geq 1$, $T > 0$, are continuous at $\hat{x}$.

PROOF. We begin by proving that $x \to \tau_1(x, G_1) \wedge T$ is continuous at $\hat{x}$. Let $G_1^{-\varepsilon}$ consist of those $a \in G_1$ whose ε-neighborhoods belong to G_1; this set is nonempty for ε small enough.

Our first observation is that, for $x \in D[0,\infty)$,

$$\lim_{\varepsilon \downarrow 0} \tau_1(x, G_1^{-\varepsilon}) = \tau_1(x, G_1). \tag{2.7}$$

Indeed, since $G_1^{-\varepsilon} \subset G_1$, we have that $\tau_1(x, G_1^{-\varepsilon}) \geq \tau_1(x, G_1)$. In particular, this proves (2.7) if $\tau_1(x, G_1) = \infty$. If $\tau_1(x, G_1) < \infty$, then, for some $\delta > 0$, $x(\tau_1(x, G_1) + \delta) \in G_1$ and, since $\bigcup_{\varepsilon > 0} G_1^{-\varepsilon} = G_1$, we have that $x(\tau_1(x, G_1) + \delta) \in G_1^{-\varepsilon}$ if ε is small enough, so $\tau_1(x, G_1) + \delta \geq \tau_1(x, G_1^{-\varepsilon})$. Since δ can be chosen arbitrarily small, we conclude that $\tau_1(x, G_1) \geq \limsup_{\varepsilon \downarrow 0} \tau_1(x, G_1^{-\varepsilon})$. The limit (2.7) is proved.

Now let $x^n = (x^n(t),\ t \geq 0)$ converge to $\hat{x}$, as $n \to \infty$. In particular, since $\hat{x}$ is continuous,

$$\lim_{n \to \infty} \sup_{t \leq T} |x^n(t) - \hat{x}(t)| = 0.$$

Let ε be arbitrary but small enough so that $G_1^{-\varepsilon}$ is nonempty, and let n be such that

$$\sup_{t \leq T} |x^n(t) - \hat{x}(t)| < \varepsilon. \tag{2.8}$$

Then, since $\hat{x} \in G_1^{-\varepsilon}$ implies $x^n \in G_1$,

$$\tau_1(x^n, G_1) \wedge T \leq \tau_1(\hat{x}, G_1^{-\varepsilon}) \wedge T$$

and, by (2.7),

$$\limsup_{n\to\infty} \tau_1(x^n, G_1) \wedge T \le \tau_1(\hat{x}, G_1) \wedge T.$$

On the other hand, under (2.8), $x^n \in G_1$ implies $\hat{x} \in G_1^\varepsilon$, and hence $\tau_1(x^n, G_1) \wedge T \ge \tau_1(\hat{x}, G_1^\varepsilon) \wedge T$, and by the conditions of the lemma,

$$\liminf_{n\to\infty} \tau_1(x^n, G_1) \wedge T \ge \tau_1(\hat{x}, G_1) \wedge T.$$

The continuity of $\tau_1(x, G_1) \wedge T$ at $\hat{x}$ is proved. Replacing G_1 by G_2 and $x(t)$ by $x(t + \tau_1(x, G_1) \wedge T)$, $t \ge 0$, we get that $x \to \zeta_1(x, G_2) \wedge T$ is continuous at $\hat{x}$.

The proof is concluded by induction if we note that, for $x \in D[0, \infty)$ and $k \ge 2$,

$$\tau_k(x, G_1) \wedge T = (\zeta_1(x, G_2) \wedge T + \tau_{k-1}(x_1, G_1) \wedge T) \wedge T,$$
$$\zeta_k(x, G_2) \wedge T = (\zeta_1(x, G_2) \wedge T + \zeta_{k-1}(x_1, G_2) \wedge T) \wedge T,$$

where

$$x_1(t) = x(t + \zeta_1(x, G_2) \wedge T). \qquad \square$$

3. The threshold queue with exceptional arrivals. In this section we prove an averaging principle for a single-server queue, called the threshold queue, which provides a critical element in our analysis of polling systems. The threshold queue is basically the standard FIFO single-server queue except that, for a given parameter $h \ge 0$, busy periods of the threshold queue begin only when the queue length first exceeds h; busy periods terminate in the normal way, whenever the queue becomes empty. We say that the server switches on when the busy periods begin and switches off when the busy periods end. Those periods during which the server is switched off are called *accumulation* periods; such a period includes the usual idle period plus a period during which arrivals are accumulating in the queue. An accumulation period and its following busy period make up a cycle.

Threshold queues correspond in the obvious way to the queues in our two-queue polling system; for example, the accumulation periods of the threshold queue representing queue 1 correspond to the busy periods of queue 2. In our general approach to the proof of the main result (cf. Theorem 2.1), the time interval $[0, T]$ is divided into subintervals sufficiently small that the total number in the system remains approximately constant during each. Then, during a subinterval, the behavior of each queue is approximated by that of a threshold queue. The main result of this section (Theorem 3.1) shows that a threshold queue also obeys an averaging principle. The averaging principle for the polling system is derived as a consequence of the averaging principles of the threshold queues defined for the subintervals.

Consider a sequence of threshold queues indexed by n. With the exception noted below, interarrival and service times form independent i.i.d. sequences, where generic interarrival and service times are denoted by ξ^n and η^n, respectively. The threshold for the nth queue is $h^n = \lfloor \sqrt{n}\, a^n \rfloor$, where a^n is a given constant. Assume that

$$\sup_n E(\xi^n)^2 < \infty, \qquad \sup_n E(\eta^n)^2 < \infty, \tag{3.1}$$

and, letting $\lambda^n = (E\xi^n)^{-1}$ and $\mu^n = (E\eta^n)^{-1}$, assume that

$$\lim_{n\to\infty} \lambda^n = \lambda > 0, \quad \lim_{n\to\infty} \mu^n = \mu > 0, \quad \lim_{n\to\infty} a^n = a > 0, \quad \lambda < \mu. \tag{3.2}$$

For technical reasons to be made clear later, we will need a slight generalization of the renewal arrival process determined by ξ^n for the nth threshold queue: within each busy period, at most one of the interarrival periods is allowed to be *exceptional*, that is, have a distribution other than that of ξ^n. We make no specific assumptions about the dependence of exceptional interarrival periods on other quantities, but we assume they are bounded, as follows.

For each $i \geq 1$, we define a nonnegative random variable $\tilde{\xi}_i^n$ and an integer-valued random variable χ_i^n that correspond to the ith cycle. Specifically, if the ith busy period has at least χ_i^n arrivals, then the χ_i^nth arrival is exceptional and the duration of the latter is taken to be $\tilde{\xi}_i^n$. There are no exceptional arrivals if the busy period has fewer than χ_i^n arrivals. We assume that there exists a family of sequences $\{\zeta_i^n(r),\ i \geq 1\}$, $r > 0$, of identically distributed nonnegative random variables such that

$$\begin{aligned} &\frac{1}{\sqrt{n}} \zeta_1^n(r) \to_P 0 \quad \text{as } n \to \infty, r > 0, \\ &\lim_{r\to\infty} \limsup_{n\to\infty} \sum_{i=1}^{\lfloor t\sqrt{n} \rfloor} P\big(\tilde{\xi}_i^n > \zeta_i^n(r)\big) = 0, \qquad t > 0, \end{aligned} \tag{3.3}$$

and that the joint distribution of the $\zeta_i^n(r)$, the normal interarrival times and the service times in the ith cycle does not depend on i. We also assume that the time of the first arrival, which we denote by $\bar{\xi}_1^n$, may have a distribution different from that of the generic interarrival time and that

$$\frac{\bar{\xi}_1^n}{\sqrt{n}} \to_P 0.$$

Introduce $X^n(t) = Q^n(nt)/\sqrt{n}$, $t \geq 0$, where $Q^n(t)$ is the queue length at t, and assume that $X^n(0) = 0$.

THEOREM 3.1. *Let $f(x)$, $x \in R_+$, denote a bounded continuous function. If conditions* (3.1)–(3.3) *hold, then for any $T > 0$,*

$$\int_0^T f(X^n(t))\,dt \to_P T\int_0^1 f(au)\,du \quad \text{as } n \to \infty.$$

PROOF. The proof consists of suitably applying the law of large numbers given by Lemma 2.4. This would be almost straightforward if the cycles were identically distributed and we could apply the identically distributed version of Lemma 2.4. However, since the threshold queue being considered does not follow exactly the same probabilistic law during each cycle (because of exceptional interarrival times), we have to apply Lemma 2.4 in all its generality, and this creates quite a few technical difficulties.

Define the times illustrated in Figure 2:

$$
\begin{aligned}
&\gamma_0^n = 0,\\
&\alpha_i^n = \inf\big(t > \gamma_{i-1}^n \colon Q^n(nt) = 1\big), \qquad i \ge 1,\\
&\beta_i^n = \inf\big(t > \gamma_{i-1}^n \colon Q^n(nt) > h^n\big), \qquad i \ge 1,\\
&\gamma_i^n = \inf\big(t > \beta_i^n \colon Q^n(nt) = 0\big), \qquad i \ge 1.
\end{aligned} \tag{3.4}
$$

Note that the β_i^n start and the γ_i^n terminate busy periods.

We prove that

$$\gamma_{\lfloor\sqrt{n}t\rfloor}^n \to_P a\left(\frac{1}{\lambda} + \frac{1}{\mu - \lambda}\right)t \quad \text{as } n \to \infty, \tag{3.5}$$

and

$$\int_0^{\gamma_{\lfloor\sqrt{n}t\rfloor}^n} f(X^n(s))\,ds \to_P t\left(\frac{1}{\lambda} + \frac{1}{\mu - \lambda}\right)\int_0^a f(u)\,du \quad \text{as } n \to \infty, \tag{3.6}$$

which immediately give the assertion of the theorem.

For $i \ge 2$, denote by $\bar{\xi}_i^n$ the time between γ_{i-1}^n and the first arrival after γ_{i-1}^n, that is, $\bar{\xi}_i^n = \alpha_i^n - \gamma_{i-1}^n$, and denote by $\{\xi_{i,k}^{n,1}, k \ge 1\}$ and $\{\xi_{i,k}^{n,2}, k \ge 1\}$ the i.i.d. sequences, with generic random variable ξ^n, from which normal interarrival times on $[\alpha_i^n, \beta_i^n]$ and $[\beta_i^n, \alpha_{i+1}^n]$, respectively, are taken. Similarly, let $\{\eta_{i,k}^n, k \ge 1\}$, $i \ge 1$, be service times on $[\beta_i^n, \gamma_i^n]$. Note that by the

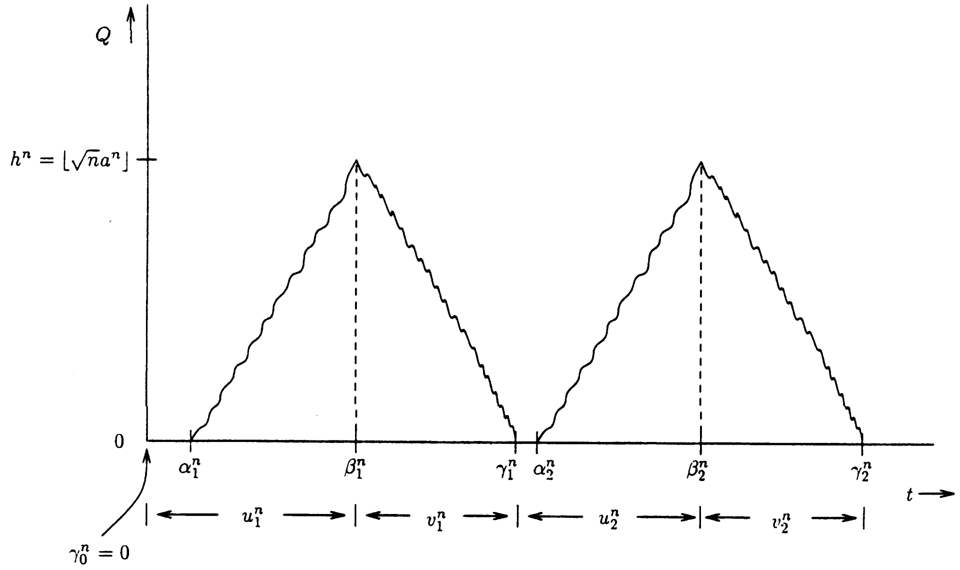

FIG. 2. *Notation for Theorem* 3.1 (*sample paths shown are rough sketches*).

conditions of the theorem, the distribution of $\{\zeta_i^n(r), \xi_{i,k}^{n,l}, \eta_{i,k}^n, l = 1,2, k \geq 1\}$ does not depend on $i = 1,2,\dots$.

Define, for $i \geq 1$,

$$A_i^n(t) = 1\left(\bar{\xi}_i^n \leq t\right) + \sum_{k=1}^{\infty} 1\left(\bar{\xi}_i^n + \sum_{j=1}^{k} \xi_{i,j}^{n,1} \leq t\right), \tag{3.7a}$$

$$\begin{aligned} B_i^n(t) = {} & \sum_{k=1}^{\chi_i^n - 1} 1\left(\sum_{j=1}^{k} \xi_{i,j}^{n,2} \leq t\right) + 1\left(\sum_{j=1}^{\chi_i^n - 1} \xi_{i,j}^{n,2} + \tilde{\xi}_i^n \leq t\right) \\ & + \sum_{k=\chi_i^n}^{\infty} 1\left(\sum_{j=1}^{k} \xi_{i,j}^{n,2} + \tilde{\xi}_i^n \leq t\right), \end{aligned} \tag{3.7b}$$

$$S_i^n(t) = \sum_{k=1}^{\infty} 1\left(\sum_{j=1}^{k} \eta_{i,j}^n \leq t\right). \tag{3.7c}$$

Note that $A_i^n = (A_i^n(t),\ t \geq 0)$ is the arrival process on $[\gamma_{i-1}^n, \beta_i^n]$, $B_i^n = (B_i^n(t), t \geq 0)$ is the arrival process on $[\beta_i^n, \gamma_i^n]$ and $S_i^n = (S_i^n(t), t \geq 0)$ is the service process on $[\beta_i^n, \gamma_i^n]$.

In a sense, the $\bar{\xi}_i^n$, $i \geq 2$, also represent exceptional interarrival times, since they are distributed differently from ξ^n. We prove that they satisfy conditions similar to those imposed on $\tilde{\xi}_i^n$.

LEMMA 3.1. *For $r > 0$, let*

$$\bar{\zeta}_i^n(r) = \max_{1 \leq k \leq \lfloor r\sqrt{n} \rfloor} \xi_{i-1,k}^{n,2}, \qquad i \geq 2.$$

Then, as $n \to \infty$,

$$\frac{\bar{\zeta}_i^n(r)}{\sqrt{n}} \to_P 0, \qquad i \geq 2,$$

and

$$\lim_{r \to \infty} \limsup_{n \to \infty} \sum_{i=2}^{\lfloor t\sqrt{n} \rfloor} P\left(\bar{\xi}_i^n > \bar{\zeta}_i^n(r)\right) = 0, \qquad t > 0.$$

PROOF. The first limit follows by (3.1). For the second, note that if $\bar{\xi}_i^n > \bar{\zeta}_i^n(r)$, then the number of arrivals in $[\beta_{i-1}^n, \gamma_i^n]$ is at least $\lfloor r\sqrt{n} \rfloor$, which can only happen if the time needed for $B_{i-1}^n(t)$ to reach $\lfloor r\sqrt{n} \rfloor$ is not greater than the time taken by $S_{i-1}^n(t)$ to become equal to $\lfloor r\sqrt{n} \rfloor + h^n + 1$. Therefore, accounting for the possibility of an exceptional arrival in $[\beta_{i-1}^n, \gamma_i^n]$, we have

$$\sum_{i=2}^{\lfloor t\sqrt{n}\rfloor} P\big(\bar{\xi}_i^n > \bar{\zeta}_i^n(r)\big) \le \sum_{i=2}^{\lfloor t\sqrt{n}\rfloor} P\left(\sum_{k=1}^{\lfloor r\sqrt{n}\rfloor - 1} \xi_{i-1,k}^{n,2} \le \sum_{k=1}^{\lfloor r\sqrt{n}\rfloor + h^n + 1} \eta_{i-1,k}^n\right)$$

$$\le t\sqrt{n}\left[P\left(\sum_{k=1}^{\lfloor r\sqrt{n}\rfloor - 1} \xi_{1,k}^{n,2} \le \frac{(\lambda^n)^{-1} + (\mu^n)^{-1}}{2} r\sqrt{n}\right)\right.$$

$$\left. + P\left(\sum_{k=1}^{\lfloor r\sqrt{n}\rfloor + h^n + 1} \eta_{1,k}^n \ge \frac{(\lambda^n)^{-1} + (\mu^n)^{-1}}{2} r\sqrt{n}\right)\right]$$

On centering the sums in the events on the right-hand side and applying Chebyshev's inequality, we get that, for r large enough,

$$\sum_{i=1}^{\lfloor t\sqrt{n}\rfloor} P\big(\bar{\xi}_i^n > \bar{\zeta}_i^n(r)\big) \le t\sqrt{n}\left[\frac{5}{\big((\lambda^n)^{-1} - (\mu^n)^{-1}\big)^2} \frac{1}{r^2 n} r\sqrt{n}\, \operatorname{Var} \xi_{1,1}^{n,2}\right.$$

$$+ \left(\frac{(\lambda^n)^{-1} - (\mu^n)^{-1}}{2} r\sqrt{n} - (\mu^n)^{-1}(h^n + 1)\right)^{-2}$$

$$\left. \times \big(r\sqrt{n} + h^n + 1\big) \operatorname{Var} \eta_{1,1}^n \right]$$

and that the latter tends to 0 as $n \to \infty$ and $r \to \infty$ by (3.1), (3.2) and $h^n = \lfloor \sqrt{n}\, a^n \rfloor$. □

For homogeneity of notation, we further set $\bar{\zeta}_1^n(r) = \bar{\xi}_1^n$. We now return to the proof of the theorem. Introduce the event

$$\Gamma^n(r) = \bigcap_{i=1}^{\lfloor t\sqrt{n}\rfloor} \big\{\tilde{\xi}_i^n \le \zeta_i^n(r),\ \bar{\xi}_i^n \le \bar{\zeta}_i^n(r)\big\}.$$

Since by (3.3) and Lemma 3.1, $P(\Gamma^n(r)) \to 1$ as $n \to \infty$ and $r \to \infty$, it is enough to prove (3.5) and (3.6) on $\Gamma^n(r)$.

Define the interval lengths (see Figure 2)

$$u_i^n = \beta_i^n - \gamma_{i-1}^n, \qquad v_i^n = \gamma_i^n - \beta_i^n, \qquad i \ge 1,$$

so that by (3.4) and (3.7),

$$\begin{aligned} u_i^n &= \inf\big(t > 0: A_i^n(nt) > h^n\big), \\ v_i^n &= \inf\big(t > 0: S_i^n(nt) - B_i^n(nt) > h^n\big) \end{aligned} \tag{3.8}$$

and

$$\gamma_i^n - \gamma_{i-1}^n = u_i^n + v_i^n. \tag{3.9}$$

Limit (3.5) on $\Gamma^n(r)$ is proved by reduction to the identically distributed case of Lemma 2.4. Since $(\gamma_i^n - \gamma_{i-1}^n,\ i \ge 1)$ are not generally identically distributed, we first construct suitable upper and lower bounds. Informally, the lower-bound process results from taking $\bar{\xi}_i^n = 0$ and $\tilde{\xi}_i^n = \zeta_i^n(r)$, and the upper-bound process results from taking $\bar{\xi}_i^n = \bar{\zeta}_i^n(r)$ and $\tilde{\xi}_i^n = 0$.

In analogy with (3.7a, b), formally define (since r is fixed, it is omitted in the new notation below)

$$\begin{aligned} \bar{A}_i^n(t) &= 1\left(\bar{\zeta}_i^n(r) \le t\right) + \sum_{k=1}^{\infty} 1\left(\bar{\zeta}_i^n(r) + \sum_{j=1}^{k} \xi_{i,j}^{n,1} \le t\right), \\ \underline{A}_i^n(t) &= 1 + \sum_{k=1}^{\infty} 1\left(\sum_{j=1}^{k} \xi_{i,j}^{n,1} \le t\right), \\ \bar{B}_i^n(t) &= 1 + \sum_{k=1}^{\infty} 1\left(\sum_{j=1}^{k} \xi_{i,j}^{n,2} \le t\right), \\ \underline{B}_i^n(t) &= \sum_{k=1}^{\infty} 1\left(\zeta_i^n(r) + \sum_{j=1}^{k} \xi_{i,j}^{n,2} \le t\right), \end{aligned} \tag{3.10}$$

and define as in (3.8) and (3.9),

$$\begin{aligned} \bar{u}_i^n &= \inf\left(t > 0 : \bar{A}_i^n(nt) > h^n\right), \\ \underline{u}_i^n &= \inf\left(t > 0 : \underline{A}_i^n(nt) > h^n\right), \\ \bar{v}_i^n &= \inf\left(t > 0 : S_i^n(nt) - \bar{B}_i^n(nt) > h^n\right), \\ \underline{v}_i^n &= \inf\left(t > 0 : S_i^n(nt) - \underline{B}_i^n(nt) > h^n\right), \end{aligned} \tag{3.11}$$

and

$$\begin{aligned} &\bar{\gamma}_i^n = \sum_{j=1}^{i} \left(\bar{u}_j^n + \bar{v}_j^n\right), \qquad \underline{\gamma}_i^n = \sum_{j=1}^{i} \left(\underline{u}_j^n + \underline{v}_j^n\right), \qquad i \ge 1, \\ &\bar{\gamma}_0^n = \underline{\gamma}_0^n = 0. \end{aligned} \tag{3.12}$$

Since $\bar{\xi}_i^n \le \bar{\zeta}_i^n(r)$, $\tilde{\xi}_i^n \le \zeta_i^n(r)$, $1 \le i \le t\sqrt{n}$, on $\Gamma^n(r)$, we have by (3.7) and (3.10) that, on $\Gamma^n(r)$,

$$\begin{aligned} \bar{A}_i^n(t) &\le A_i^n(t) \le \underline{A}_i^n(t), \\ \underline{B}_i^n(t) &\le B_i^n(t) \le \bar{B}_i^n(t), \end{aligned} \tag{3.13}$$

and hence by (3.8) and (3.11), for $1 \le i \le t\sqrt{n}$,

$$\underline{u}_i^n \le u_i^n \le \bar{u}_i^n, \qquad \underline{v}_i^n \le v_i^n \le \bar{v}_i^n \quad \text{on } \Gamma^n(r), \tag{3.14}$$

and then by (3.9) and (3.12), for $1 \le i \le t\sqrt{n}$,

$$\underline{\gamma}_i^n - \underline{\gamma}_{i-1}^n \le \gamma_i^n - \gamma_{i-1}^n \le \bar{\gamma}_i^n - \bar{\gamma}_{i-1}^n \quad \text{on } \Gamma^n(r). \tag{3.15}$$

Now we prove (3.5) for $\bar{\gamma}_{\lfloor\sqrt{n}t\rfloor}^n$ and $\underline{\gamma}_{\lfloor\sqrt{n}t\rfloor}^n$; this will imply (3.5) for $\gamma_{\lfloor\sqrt{n}t\rfloor}^n$ on $\Gamma^n(r)$. Consider the lower-bound process. The proof for $\bar{\gamma}_{\lfloor\sqrt{n}t\rfloor}^n$ is similar.

First note that, by (3.10) and (3.7c),

$$\underline{A}_i^n(t) = \inf\left(k\colon \sum_{j=1}^{k} \xi_{i,j}^{n,1} > t\right),$$

$$\underline{B}_i^n(t) = \inf\left(k\colon \zeta_i^n(r) + \sum_{j=1}^{k+1} \xi_{i,j}^{n,2} > t\right),$$

$$S_i^n(t) = \inf\left(k\colon \sum_{j=1}^{k+1} \eta_{i,j}^n > t\right).$$

Since $\{\xi_{i,k}^{n,l},\ k \geq 1\}$, $l = 1,2$, are i.i.d., we have by (3.1) and (3.2) that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor \sqrt{n}\, t \rfloor} \xi_{i,k}^{n,l} \to_P \frac{t}{\lambda}, \qquad l = 1,2, \qquad \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor \sqrt{n}\, t \rfloor} \eta_{i,k}^n \to_P \frac{t}{\mu},$$

and hence, with the use of the first relation in (3.3), Lemma 2.1 and (3.11) yield

$$\begin{gathered} \frac{1}{\sqrt{n}} \underline{A}_i^n(\sqrt{n}\, t) \to_P \lambda t, \qquad \frac{1}{\sqrt{n}} \underline{B}_i^n(\sqrt{n}\, t) \to_P \lambda t, \\ \frac{1}{\sqrt{n}} S_i^n(\sqrt{n}\, t) \to_P \mu t, \qquad \sqrt{n}\, \underline{u}_i^n \to_P \frac{a}{\lambda}, \\ \sqrt{n}\, \underline{v}_i^n \to_P \frac{a}{\mu - \lambda}, \qquad i \geq 1, \end{gathered} \tag{3.16}$$

and then, by (3.12),

$$\sqrt{n}\left(\underline{\gamma}_i^n - \underline{\gamma}_{i-1}^n\right) \to_P a\left(\frac{1}{\lambda} + \frac{1}{\mu - \lambda}\right), \qquad i \geq 1. \tag{3.17}$$

Since

$$\underline{\gamma}_{\lfloor \sqrt{n}\, t \rfloor}^n = \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor \sqrt{n}\, t \rfloor} \sqrt{n}\left(\underline{\gamma}_k^n - \underline{\gamma}_{k-1}^n\right)$$

and since $(\underline{\gamma}_i^n - \underline{\gamma}_{i-1}^n,\ i \geq 1)$ are identically distributed by construction, we would have, in view of Lemma 2.4,

$$\underline{\gamma}_{\lfloor \sqrt{n}\, t \rfloor}^n \to_P a\left(\frac{1}{\lambda} + \frac{1}{\mu - \lambda}\right) t \tag{3.18}$$

provided

$$\lim_{k \to \infty} \limsup_{n \to \infty} \sqrt{n}\, P\left(\sqrt{n}\left(\underline{\gamma}_1^n - \underline{\gamma}_0^n\right) > k\right) = 0. \tag{3.19}$$

By (3.12), this would follow from

$$\begin{gathered} \lim_{k \to \infty} \limsup_{n \to \infty} \sqrt{n}\, P\left(\sqrt{n}\, \underline{u}_1^n > k\right) = 0, \\ \lim_{k \to \infty} \limsup_{n \to \infty} \sqrt{n}\, P\left(\sqrt{n}\, \underline{v}_1^n > k\right) = 0. \end{gathered} \tag{3.20}$$

We prove only the second of these limits; the proof of the other is similar and somewhat easier. By (3.11) we have

$$\begin{aligned}
P\big(\sqrt{n}\,\underline{v}_1^n > k\big) &= P\Big(\sup_{t\le k}\big(S_1^n(\sqrt{n}\,t) - \underline{B}_1^n(\sqrt{n}\,t)\big) \le h^n\Big)\\
&\le P\big(S_1^n(\sqrt{n}\,k) - \underline{B}_1^n(\sqrt{n}\,k) \le h^n\big)\\
&\le P\big(\underline{B}_1^n(\sqrt{n}\,k) > \tfrac{1}{2}(\lambda^n + \mu^n)\sqrt{n}\,k\big)\\
&\quad + P\big(S_1^n(\sqrt{n}\,k) \le h^n + \tfrac{1}{2}(\lambda^n + \mu^n)\sqrt{n}\,k\big).
\end{aligned} \tag{3.21}$$

Since by (3.10),

$$P\big(\underline{B}_1^n(\sqrt{n}\,k) > \tfrac{1}{2}(\lambda^n + \mu^n)\sqrt{n}\,k\big) \le P\left(\sum_{j=1}^{\lfloor((\lambda^n+\mu^n)/2)\sqrt{n}\,k\rfloor} \xi_{1,j}^{n,2} \le \sqrt{n}\,k\right),$$

and by (3.7c),

$$P\big(S_1^n(\sqrt{n}\,k) \le h^n + \tfrac{1}{2}(\lambda^n + \mu^n)\sqrt{n}\,k\big) \le P\left(\sum_{j=1}^{\lfloor((\lambda^n+\mu^n)/2)\sqrt{n}\,k\rfloor + h^n + 1} \eta_{1,j}^n > \sqrt{n}\,k\right),$$

we have by (3.21), in analogy with the proof of Lemma 3.1, that for some $C > 0$, $k > 4a/(\mu - \lambda)$ and n large enough, $\sqrt{n}\,P(\sqrt{n}\,\underline{v}_1^n > k) \le C/k$, thus proving (3.20) and hence (3.18). This ends the proof of (3.5). Note that the proof for $\bar{\gamma}^n_{\lfloor\sqrt{n}\,t\rfloor}$ also invokes Lemma 3.1 to get analogues of (3.16).

To prove (3.6) on $\Gamma^n(r)$, we apply the part of Lemma 2.4 dealing with nonidentically distributed summands. That is, we prove that

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor\sqrt{n}\,t\rfloor} P\Bigg(\Bigg\{\Bigg|&\sqrt{n}\int_{\gamma_{i-1}^n}^{\gamma_i^n} f(X^n(s))\,ds\\
&-\left(\frac{1}{\lambda} + \frac{1}{\mu-\lambda}\right)\\
&\times\int_0^a f(u)\,du\Bigg| > \varepsilon\Bigg\} \cap \Gamma^n(r)\Bigg) = 0 \quad \text{if } \varepsilon > 0,
\end{aligned} \tag{3.22}$$

and

$$\lim_{k\to\infty}\limsup_{n\to\infty} \sum_{i=1}^{\lfloor\sqrt{n}\,t\rfloor} P\left(\left\{\sqrt{n}\left|\int_{\gamma_{i-1}^n}^{\gamma_i^n} f(X^n(s))\,ds\right| > k\right\} \cap \Gamma^n(r)\right) = 0. \tag{3.23}$$

Note that (3.23) is easy: by the right inequality in (3.15) and the boundedness of f, we have, letting $\|\cdot\|$ denote the sup norm,

$$\begin{aligned}
\limsup_{n\to\infty} &\sum_{i=1}^{\lfloor\sqrt{n}\,t\rfloor} P\left(\left\{\sqrt{n}\left|\int_{\gamma_{i-1}^n}^{\gamma_i^n} f(X^n(s))\,ds\right| > k\right\} \cap \Gamma^n(r)\right)\\
&\le \limsup_{n\to\infty} \sum_{i=1}^{\lfloor\sqrt{n}\,t\rfloor} P\big(\sqrt{n}\,(\bar{\gamma}_i^n - \bar{\gamma}_{i-1}^n)\|f\| > k\big),
\end{aligned}$$

which tends to 0 as $k \to \infty$ by an analogue of (3.19) for $\bar{\gamma}_i^n$ and by the fact that the $(\bar{\gamma}_i^n - \bar{\gamma}_{i-1}^n)$, $i \ge 2$, are identically distributed.

By (3.9), (3.22) would follow if

$$\lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor\sqrt{n}\,t\rfloor}P\left(\left\{\left|\sqrt{n}\int_0^{u_i^n}f(X^n(s+\gamma_{i-1}^n))\,ds\right.\right.\right.$$
$$\left.\left.\left.-\frac{1}{\lambda}\int_0^a f(u)\,du\right|>\varepsilon\right\}\cap\Gamma^n(r)\right)=0$$

and

$$\lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor\sqrt{n}\,t\rfloor}P\left(\left\{\left|\sqrt{n}\int_0^{v_i^n}f(X^n(s+\beta_i^n))\,ds\right.\right.\right.$$
$$\left.\left.\left.-\frac{1}{\mu-\lambda}\int_0^a f(u)\,du\right|>\varepsilon\right\}\cap\Gamma^n(r)\right)=0. \tag{3.24}$$

These limits have similar proofs; we prove only (3.24).

First, by the second set of inequalities in (3.14) and the fact that $\{\bar{v}_i^n,\ i\geq 1\}$ and $\{\underline{v}_i^n,\ i\geq 1\}$ each consist of identically distributed random variables, for $\delta>0$, $1\leq i\leq t\sqrt{n}$,

$$P\left(\left\{\left|\sqrt{n}\,v_i^n-\frac{a}{\mu-\lambda}\right|>\delta\right\}\cap\Gamma^n(r)\right)$$
$$\leq P\left(\left|\sqrt{n}\,\bar{v}_1^n-\frac{a}{\mu-\lambda}\right|>\delta\right)+P\left(\left|\sqrt{n}\,\underline{v}_1^n-\frac{a}{\mu-\lambda}\right|>\delta\right)$$

and hence

$$\limsup_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor\sqrt{n}\,t\rfloor}P\left(\left\{\left|\sqrt{n}\,v_i^n-\frac{a}{\mu-\lambda}\right|>\delta\right\}\cap\Gamma^n(r)\right)$$
$$\leq t\limsup_{n\to\infty}P\left(\left|\sqrt{n}\,\bar{v}_1^n-\frac{a}{\mu-\lambda}\right|>\delta\right)$$
$$+\,t\limsup_{n\to\infty}P\left(\left|\sqrt{n}\,\underline{v}_1^n-\frac{a}{\mu-\lambda}\right|>\delta\right)$$
$$=0, \tag{3.25}$$

where the last equality follows since $\sqrt{n}\,\underline{v}_1^n\to_P a/(\mu-\lambda)$ [see (3.16)] and $\sqrt{n}\,\bar{v}_1^n\to_P a/(\mu-\lambda)$. [The latter is proved analogously to (3.16).] Next,

$$P\left(\left\{\left|\sqrt{n}\int_0^{v_i^n}f(X^n(s+\beta_i^n))\,ds-\frac{1}{\mu-\lambda}\int_0^a f(u)\,du\right|>\varepsilon\right\}\cap\Gamma^n(r)\right)$$
$$\leq P\left(\left\{\int_0^{\sqrt{n}\,v_i^n\wedge a/(\mu-\lambda)}\left|f\left(X^n\left(\frac{u}{\sqrt{n}}+\beta_i^n\right)\right)\right.\right.\right.$$

$$-f(a-(\mu-\lambda)u)\biggr|\,du>\frac{\varepsilon}{2}\biggr\}\cap\Gamma^n(r)\biggr)$$

$$+P\biggl(\biggl\{\|f\|\cdot\biggl|\sqrt{n}\,v_i^n-\frac{a}{\mu-\lambda}\biggr|>\frac{\varepsilon}{2}\biggr\}\cap\Gamma^n(r)\biggr).$$

By (3.25), $1/\sqrt{n}$ times the sum from 1 to $\lfloor t\sqrt{n}\rfloor$ of the second term on the right tends to 0 in probability as $n\to\infty$, so the proof of (3.24) will be finished by proving

$$\lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor\sqrt{n}\,t\rfloor}P\biggl(\biggl\{\int_0^{\sqrt{n}\,v_i^n\wedge a/(\mu-\lambda)}\biggl|f\biggl(X^n\biggl(\frac{u}{\sqrt{n}}+\beta_i^n\biggr)\biggr)-f(a-(\mu-\lambda)u)\biggr|\,du>\frac{\varepsilon}{2}\biggr\}\cap\Gamma^n(r)\biggr)=0. \tag{3.26}$$

We prove first that, for $\eta>0$,

$$\limsup_{n}\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor\sqrt{n}\,t\rfloor}P\biggl(\biggl\{\sup_{u\le\sqrt{n}\,v_i^n\wedge a/(\mu-\lambda)}\biggl|X^n\biggl(\frac{u}{\sqrt{n}}+\beta_i^n\biggr)-(a-(\mu-\lambda)u)\biggr|>\eta\biggr\}\cap\Gamma^n(r)\biggr)=0. \tag{3.27}$$

By construction,

$$X^n(u+\beta_i^n)=\frac{h^n+1}{\sqrt{n}}+\frac{B_i^n(nu)-S_i^n(nu)}{\sqrt{n}},\qquad u\in[0,v_i^n],$$

and so, since $h^n=\lfloor\sqrt{n}\,a^n\rfloor$,

$$\begin{aligned}&P\biggl(\biggl\{\sup_{u\le\sqrt{n}\,v_i^n\wedge a/(\mu-\lambda)}\biggl|X^n\biggl(\frac{u}{\sqrt{n}}+\beta_i^n\biggr)-(a-(\mu-\lambda)u)\biggr|>\eta\biggr\}\cap\Gamma^n(r)\biggr)\\&\quad\le P\biggl(\biggl\{\sup_{u\le a/(\mu-\lambda)}\biggl|\frac{B_i^n(\sqrt{n}\,u)}{\sqrt{n}}-\lambda u\biggr|>\frac{\eta}{3}\biggr\}\cap\Gamma^n(r)\biggr)\\&\qquad+P\biggl(\sup_{u\le a/(\mu-\lambda)}\biggl|\frac{S_i^n(\sqrt{n}\,u)}{\sqrt{n}}-\mu u\biggr|>\frac{\eta}{3}\biggr)\\&\qquad+1\biggl(\biggl|\frac{\lfloor\sqrt{n}\,a^n\rfloor+1}{\sqrt{n}}-a\biggr|>\frac{\eta}{3}\biggr).\end{aligned}$$

Since the distributions of $(\underline{B}_i^n(t),\ t\ge 0)$, $(\overline{B}_i^n(t),\ t\ge 0)$ and $(S_i^n(t),\ t\ge 0)$ do not depend on i, we conclude from (3.2) and (3.13) that the left-hand side of

(3.27) is not greater than

$$t \limsup_{n\to\infty} P\left(\sup_{u \le a/(\mu-\lambda)} \left| \frac{1}{\sqrt{n}} \overline{B}_1^n(\sqrt{n}\,u) - \lambda u \right| > \frac{\eta}{3} \right)$$
$$+ t \limsup_{n\to\infty} P\left(\sup_{u \le a/(\mu-\lambda)} \left| \frac{1}{\sqrt{n}} \underline{B}_1^n(\sqrt{n}\,u) - \lambda u \right| > \frac{\eta}{3} \right)$$
$$+ t \limsup_{n\to\infty} P\left(\sup_{u \le a/(\mu-\lambda)} \left| \frac{S_1^n(\sqrt{n}\,u)}{\sqrt{n}} - \mu u \right| > \frac{\eta}{3} \right),$$

which is zero by (3.16) (the same relation obviously holds for both $\overline{B}_1^n$ and $\underline{B}_1^n$); (3.27) is proved.

Now on the event

$$\left\{ \sup_{u \le \sqrt{n}\, v_i^n \wedge a/(\mu-\lambda)} \left| X^n\left(\frac{u}{\sqrt{n}} + \beta_i^n \right) - (a - (\mu - \lambda) u) \right| \le \eta \right\},$$

we have that $X^n(u/\sqrt{n} + \beta_i^n) \le a + \eta$, $u \in [0, \sqrt{n}\, v_i^n \wedge a/(\mu - \lambda)]$ and, therefore, for $u \in [0, \sqrt{n}\, v_i^n \wedge a/(\mu - \lambda)]$,

$$\left| f\left(X^n\left(\frac{u}{\sqrt{n}} + \beta_i^n \right) \right) - f(a - (\mu - \lambda) u) \right| \le \omega_f(\eta, a + \eta),$$

where $\omega_f(\delta, T)$ is the modulus of continuity of f on $[0, T]$ for partitions of diameter δ. This implies by the continuity of f that, for all η small enough and for all i,

$$\left\{ \sup_{u \le \sqrt{n}\, v_i^n \wedge a/(\mu-\lambda)} \left| X^n\left(\frac{u}{\sqrt{n}} + \beta_i^n \right) - (a - (\mu - \lambda) u) \right| \le \eta \right\}$$
$$\subset \left\{ \int_0^{\sqrt{n}\, v_i^n \wedge a/(\mu-\lambda)} \left| f\left(X^n\left(\frac{u}{\sqrt{n}} + \beta_i^n \right) \right) - f(a - (\mu - \lambda) u) \right| du \le \frac{\varepsilon}{2} \right\},$$

so for η small enough

$$P\left(\left\{ \int_0^{\sqrt{n}\, v_i^n \wedge a/(\mu-\lambda)} \left| f\left(X^n\left(\frac{u}{\sqrt{n}} + \beta_i^n \right) \right) \right.\right.\right.$$
$$\left.\left.\left. - f(a - (\mu - \lambda) u) \right| du > \frac{\varepsilon}{2} \right\} \cap \Gamma^n(r) \right)$$
$$\le P\left(\left\{ \sup_{u \le \sqrt{n}\, v_i^n \wedge a/(\mu-\lambda)} \left| X^n\left(\frac{u}{\sqrt{n}} + \beta_i^n \right) \right.\right.\right.$$
$$\left.\left.\left. - (a - (\mu - \lambda) u) \right| > \eta \right\} \cap \Gamma^n(r) \right)$$

and (3.26) follows by (3.27). Thus (3.24), (3.22) and (3.6) are proved. This completes the proof of the theorem. □

REMARK. As can be seen from the proof, it is not necessary that two interarrival times be independent when one is taken from an accumulation and the other from a busy period. The only thing that matters is that interarrival times within each busy period and each accumulation period be independent (except, perhaps, for exceptional interarrival times). The theorem still holds under this milder condition. This observation will be exploited in the proof of Theorem 2.1.

The limiting behavior of virtual waiting times $W^n(t)$ in the threshold queue is given as follows. [Note that $W^n(t)$ is not the unfinished work at time t unless the server is switched on at that time.]

THEOREM 3.2. *Under the conditions of Theorem* 3.1,

$$\int_0^T f\left(\frac{W^n(nt)}{\sqrt{n}}\right) dt \to_P T\int_0^1 f\left(\frac{au}{\lambda}\right) du.$$

PROOF. For a proof similar to that of Theorem 3.1, one uses the relations

$$\begin{aligned} W^n(t + n\gamma_i^n) &= (nu_i^n - t) + \inf(u\colon S_i^n(u) \ge A_i^n(t)), \qquad 0 \le t \le nu_i^n,\\ W^n(t + n\beta_i^n) &= \inf(u\colon S_i^n(u) \ge h^n + 1 + B_i^n(t)) - t, \qquad 0 \le t \le nv_i^n. \end{aligned}$$

By the above and Lemma 2.1,

$$\begin{aligned} \frac{W^n(\sqrt{n}\,t + n\gamma_i^n)}{\sqrt{n}} &\to_P \frac{a}{\lambda} - t + \frac{\lambda}{\mu}t, \qquad 0 \le t \le \frac{a}{\lambda},\\ \frac{W^n(\sqrt{n}\,t + n\beta_i^n)}{\sqrt{n}} &\to_P \frac{\lambda t + a}{\mu} - t, \qquad 0 \le t \le \frac{a}{\mu - \lambda}. \end{aligned}$$

Continuing as in the proof of Theorem 3.1, we get

$$\int_0^T f\left(\frac{W^n(nt)}{\sqrt{n}}\right) dt \to_P T\int_0^1 \left[\left(1 - \frac{\lambda}{\mu}\right) f\left(a\left(\frac{1-u}{\lambda} + \frac{u}{\mu}\right)\right) + \left(\frac{\lambda}{\mu}\right) f\left(\frac{au}{\mu}\right)\right] du.$$

Changes of variables on the right-hand side then yield the theorem. □

4. Proof of Theorem 2.1.

For convenience, we consider the first queue ($l = 1$) throughout. Also, we assume initially that f is bounded and nonnegative. The general case will be handled by a localization argument after the result is shown for bounded f. Our first observation is that it is sufficient to prove that, for all δ, K, with $0 < \delta < K$,

$$\begin{aligned} &\int_0^T f(X_1^n(t)) \cdot 1(\delta \le X^n(t) \le K)\, dt\\ &\qquad \to_d \int_0^T \left(\int_0^1 f(uX(t))\, du\right) \cdot 1(\delta \le X(t) \le K)\, dt. \end{aligned} \tag{4.1}$$

To see this, note that

$$\begin{aligned}(4.2)\qquad &\int_0^T [1(0 \le X^n(t) < \delta) + 1(X^n(t) > K)]\,dt\\ &\quad \to_d \int_0^T [1(0 \le X(t) < \delta) + 1(X(t) > K)]\,dt,\end{aligned}$$

since $X^n \to_d X$ and since the right-hand side of (4.2) is continuous in $D[0,\infty)$ almost everywhere with respect to the measure induced by RBM. Since

$$\begin{aligned}&\left| \int_0^T f(X_1^n(t)) \cdot 1(\delta \le X^n(t) \le K)\,dt - \int_0^T f(X_1^n(t))\,dt \right|\\ &\quad \le \|f\| \left[\int_0^T [1(0 \le X^n(t) < \delta) + 1(X^n(t) > K)]\,dt \right],\end{aligned}$$

we thus obtain, for any constant $\eta > 0$,

$$\begin{aligned}(4.3)\qquad &\lim_{\substack{\delta\to 0\\ K\to\infty}} \limsup_{n\to\infty} P\left(\left| \int_0^T f(X_1^n(t)) \cdot 1(\delta \le X^n(t) \le K)\,dt \right.\right.\\ &\qquad\qquad \left.\left. - \int_0^T f(X_1^n(t))\,dt \right| > \eta \right) = 0\end{aligned}$$

and

$$\begin{aligned}(4.4)\qquad &\lim_{\substack{\delta\to 0\\ K\to\infty}} P\left(\left| \left| \int_0^T \left(\int_0^1 f(uX(t))\,du \right) dt - \int_0^T \left(\int_0^1 f(uX(t))\,du \right) \right.\right.\right.\\ &\qquad\qquad \left.\left.\left. \cdot 1(\delta \le X(t) \le K)\,dt \right| \right| > \eta \right) = 0.\end{aligned}$$

By Theorem 4.2 in Billingsley (1968), the assertion of Theorem 2.1 follows from (4.1), (4.3) and (4.4).

It remains to prove (4.1). We follow the approach outlined at the beginning of Section 3. The heart of the argument below uses the threshold queue of Section 3 to construct bounds for individual queue lengths.

Let ε, $0 < \varepsilon < \delta/2$, be such that $N = (K - \delta)/\varepsilon$ is an integer and let $r(\varepsilon) < \varepsilon/2$. We specify $r(\varepsilon)$ later in Lemma 4.1. Let $a_i(\varepsilon) = \delta + i\varepsilon$, $0 \le i \le N$, and denote, for $0 \le i \le N$,

$$B_{r(\varepsilon)}(\varepsilon, i) = (a_i(\varepsilon) - r(\varepsilon), a_i(\varepsilon) + r(\varepsilon)),$$

$$C_{r(\varepsilon)}(\varepsilon, i) = (0, a_i(\varepsilon) - \varepsilon + r(\varepsilon)) \cup (a_i(\varepsilon) + \varepsilon - r(\varepsilon), \infty).$$

Introduce the times

$$\begin{aligned}&\zeta_0^n(\varepsilon, i) = 0, \qquad 0 \le i \le N,\\ (4.5)\qquad &\tau_k^n(\varepsilon, i) = \inf(t > \zeta_{k-1}^n(\varepsilon, i) \colon X^n(t) \in B_{r(\varepsilon)}(\varepsilon, i)),\\ &\qquad\qquad\qquad k \ge 1, 0 \le i \le N,\\ &\zeta_k^n(\varepsilon, i) = \inf(t > \tau_k^n(\varepsilon, i) \colon X^n(t) \in C_{r(\varepsilon)}(\varepsilon, i)),\\ &\qquad\qquad\qquad k \ge 1, 0 \le i \le N,\end{aligned}$$

and for the limit process

$$\begin{aligned} &\zeta_0(\varepsilon,i) = 0, \qquad 0 \le i \le N,\\ &\tau_k(\varepsilon,i) = \inf\bigl(t > \zeta_{k-1}(\varepsilon,i)\colon X(t) \in B_{r(\varepsilon)}(\varepsilon,i)\bigr),\\ (4.6)\qquad &\hphantom{\tau_k(\varepsilon,i) = \inf\bigl(t > \zeta_{k-1}(\varepsilon,i)\colon X(t)}\qquad k \ge 1, 0 \le i \le N,\\ &\zeta_k(\varepsilon,i) = \inf\bigl(t > \tau_k(\varepsilon,i)\colon X(t) \in C_{r(\varepsilon)}(\varepsilon,i)\bigr),\\ &\hphantom{\tau_k(\varepsilon,i) = \inf\bigl(t > \zeta_{k-1}(\varepsilon,i)\colon X(t)}\qquad k \ge 1, 0 \le i \le N. \end{aligned}$$

Note that

$$(4.7)\quad [\tau_k^n(\varepsilon,i), \zeta_k^n\,(\varepsilon,i)) \cap [\tau_{k'}^n(\varepsilon,i'), \zeta_{k'}^n\,(\varepsilon,i')) = \varnothing, \quad (k,i) \neq (k',i'),$$

and

$$(4.8)\quad \begin{aligned} &\left| 1(\delta \le X^n(t) \le K) - \sum_{k=1}^{\infty} \sum_{i=0}^{N} 1\bigl(t \in \bigl[\tau_k^n(\varepsilon,i), \zeta_k^n(\varepsilon,i)\bigr)\bigr) \right|\\ &\qquad \le 1(\delta - \varepsilon \le X^n(t) \le \delta) + 1(K \le X^n(t) \le K + \varepsilon), \end{aligned}$$

and that these properties hold for the limit process $X(t)$ as well, that is, with $\tau_k^n(\varepsilon,i)$ and $\zeta_k^n(\varepsilon,i)$ replaced by $\tau_k(\varepsilon,i)$ and $\zeta_k(\varepsilon,i)$ in (4.7) and (4.8).

LEMMA 4.1. (i) *With probability* 1,

$$\tau_k(\varepsilon,i) < \zeta_k(\varepsilon,i) \quad \text{on } \{\tau_k(\varepsilon,i) < \infty\},$$

and

$$\lim_{k\to\infty} P\Bigl(\min_{0\le i\le N} \zeta_k(\varepsilon,i) \le T\Bigr) = 0.$$

(ii) *The parameter $r(\varepsilon)$ can be chosen so that*

$$\begin{aligned} &\bigl(X^n, (\tau_k^n(\varepsilon,i) \wedge T, \zeta_k^n(\varepsilon,i) \wedge T)_{k\ge 1, 0\le i\le N}\bigr)\\ &\qquad \to_d \bigl(X, (\tau_k(\varepsilon,i) \wedge T, \zeta_k(\varepsilon,i) \wedge T)_{k\ge 1, 0\le i\le N}\bigr), \end{aligned}$$

where convergence is in $D[0,\infty) \times R^\infty$.

PROOF. The first part follows by the continuity of X.

For the second part, note that, in the notation of Lemma 2.5,

$$\begin{aligned} \tau_k^n(\varepsilon,i) &= \tau_k\bigl(X^n, B_{r(\varepsilon)}(\varepsilon,i)\bigr), & \zeta_k^n(\varepsilon,i) &= \zeta_k\bigl(X^n, C_{r(\varepsilon)}(\varepsilon,i)\bigr),\\ \tau_k(\varepsilon,i) &= \tau_k\bigl(X, B_{r(\varepsilon)}(\varepsilon,i)\bigr), & \zeta_k(\varepsilon,i) &= \zeta_k\bigl(X, C_{r(\varepsilon)}(\varepsilon,i)\bigr). \end{aligned}$$

Therefore, by the continuous mapping theorem, since the X^n converge in distribution to X, the desired result would follow if the maps $x \to \tau_k(x, B_{r(\varepsilon)}(\varepsilon,i)) \wedge T$ and $x \to \zeta_k(x, C_{r(\varepsilon)}(\varepsilon,i)) \wedge T$, $x \in D[0,\infty)$, were continuous almost surely with respect to the distribution of X. Since X is continu-

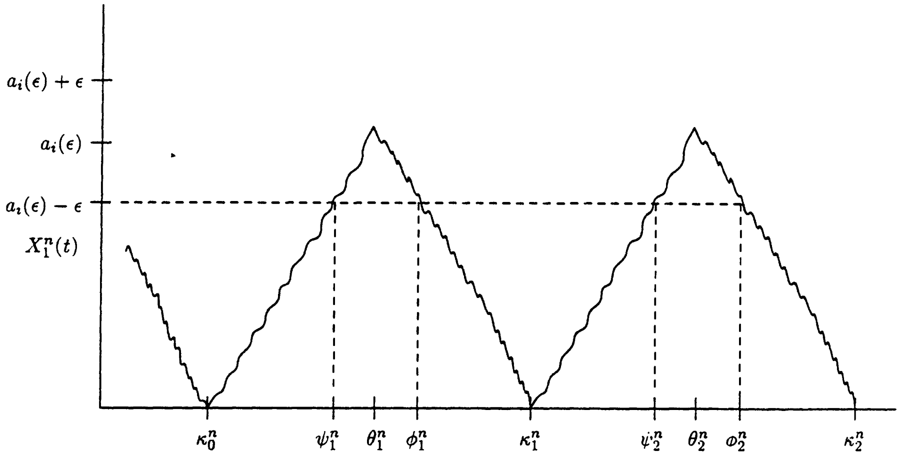

FIG. 3. *First passage times.*

ous, by Lemma 2.5 this is implied by

$$P\Big(\lim_{\eta \downarrow 0} \tau_k\big(X, B_{r(\varepsilon)+\eta}(\varepsilon, i)\big) \wedge T = \tau_k\big(X, B_{r(\varepsilon)}(\varepsilon, i)\big) \wedge T\Big) = 1,$$

$$P\Big(\lim_{\eta \downarrow 0} \zeta_k\big(X, C_{r(\varepsilon)+\eta}(\varepsilon, i)\big) \wedge T = \zeta_k\big(X, C_{r(\varepsilon)}(\varepsilon, i)\big) \wedge T\Big) = 1,$$

and the existence of $r(\varepsilon)$ satisfying the latter follows by the fact that if $\xi(r)$, $r \geq 0$, is an increasing process, then the number of different r's such that $P(\Delta\xi(r) > 0) > 0$, where $\Delta\xi(r)$ is the jump of ξ at r, is at most countable; the argument is standard [see, e.g., Billingsley (1968), Section 15, and Jacod and Shiryaev (1987), Section VI.3.12]. □

In the sequel, we assume that $r(\varepsilon)$ is chosen as required by Lemma 4.1. Our next step is to construct approximations of X_1^n on each interval $[\tau_k^n(\varepsilon, i), \zeta_k^n(\varepsilon, i))$ that are derived from the threshold queue of Section 3; one will serve as a lower bound and one as an upper bound. Fixing i, k and ε, we define the first passage times (see Figure 3):

$$\begin{aligned}
\kappa_0^n &= \inf\big(t > \tau_k^n(\varepsilon, i) \colon X_1^n(t) = 0\big),\\
\theta_j^n &= \inf\big(t > \kappa_{j-1}^n \colon X_2^n(t) = 0\big), \qquad j \geq 1,\\
\kappa_j^n &= \inf\big(t > \theta_j^n \colon X_1^n(t) = 0\big), \qquad j \geq 1, \qquad (4.9)\\
\psi_j^n &= \inf\big(t > \kappa_{j-1}^n \colon X_1^n(t) > a_i(\varepsilon) - \varepsilon\big), \qquad j \geq 1,\\
\phi_j^n &= \inf\big(t > \theta_j^n \vee \psi_j^n \colon X_1^n(t) \leq a_i(\varepsilon) - \varepsilon + 1/\sqrt{n}\big), \qquad j \geq 1.
\end{aligned}$$

Note that $n\kappa_j^n$ is a service completion time for the first queue, $n\theta_j^n$ is a service completion time for the second queue and if $\kappa_j^n \le \zeta_k^n(\varepsilon, i) < \infty$, then $n\psi_j^n$ is an arrival time for the first queue, $n\phi_j^n$ is a service completion time for the first queue and $X_1^n(\psi_j^n) = X_1^n(\phi_j^n) = [\lfloor \sqrt{n}(a_i(\varepsilon) - \varepsilon)\rfloor + 1]/\sqrt{n}$. We also set

$$\nu^n = \min\big(j\colon \kappa_{j+1}^n > \zeta_k^n(\varepsilon, i) \wedge T\big);$$
$$\nu^n = 0 \quad \text{if } \kappa_0^n > \zeta_k^n(\varepsilon, i) \wedge T. \tag{4.10}$$

Let the arrivals on $[n\kappa_0^n, \infty)$ be numbered successively starting from 1. Let $\tilde{\xi}_1^n$ denote the time period between $n\kappa_0^n$ and the first arrival. Denote by $\tilde{\xi}_l^n$, $l \ge 2$, the times between the $(l-1)$st and lth of these arrivals. Obviously, $\{\tilde{\xi}_l^n,\ l \ge 2\}$ is a set of i.i.d. random variables with the distribution of the generic interarrival time for the first queue.

Let $\tilde{\chi}_j^{n,1}$ be the index of the arrival occurring at or just before $n\psi_j^n$, $j \ge 1$, and let $\tilde{\chi}_j^{n,2}$ be the index of the arrival occurring at or just after $n\phi_j^n$, $j \ge 1$. For $j \ge 1$, let v_j^n be the time period between $n\phi_j^n$ and the $\tilde{\chi}_j^{n,2}$th arrival. Denote by $\{\tilde{\eta}_{j,l}^n,\ l \ge 1\}$, $j = 1, 2, \dots$, independent copies of the sequence of service times, which are also independent of $\{\tilde{\xi}_l^n,\ l \ge 1\}$. Again by the i.i.d. assumptions, we may assume that, for each $1 \le j \le v^n$, the service times for completions in $(\phi_j^n, \kappa_j^n]$ are $\tilde{\eta}_{j,1}^n, \tilde{\eta}_{j,2}^n, \dots$.

Now consider the threshold queue with the threshold $h^n = \lfloor \sqrt{n}(a_i(\varepsilon) - \varepsilon)\rfloor$ which has the sequence

$$\left\{\tilde{\xi}_1^n, \tilde{\xi}_2^n, \dots, \tilde{\xi}_{\tilde{\chi}_1^{n,1}}^n, v_1^n, \tilde{\xi}_{\tilde{\chi}_1^{n,2}+1}^n, \dots, \tilde{\xi}_{\tilde{\chi}_2^{n,1}}^n, v_2^n, \tilde{\xi}_{\tilde{\chi}_2^{n,2}+1}^n, \dots\right\}$$

of interarrival times; service times in the jth busy period of this queue are $\tilde{\eta}_{j,l}^n$, $l = 1, 2, \dots$. Denote by $\tilde{X}_1^n(t)$ the length of this queue, normalized and scaled as in Theorem 3.1. Also let $\tilde{\beta}_i^n$ and $\tilde{\gamma}_i^n$ be defined for this queue as in (3.4).

Then the construction above yields

$$\tilde{X}_1^n(t) = \begin{cases} X_1^n\big(t - \tilde{\gamma}_{j-1}^n + \kappa_{j-1}^n\big), & t \in \big[\tilde{\gamma}_{j-1}^n, \tilde{\beta}_j^n\big], 1 \le j \le \nu^n, \\ X_1^n\big(t - \tilde{\beta}_j^n + \phi_j^n\big), & t \in \big[\tilde{\beta}_j^n, \tilde{\gamma}_j^n\big], 1 \le j \le \nu^n, \end{cases} \tag{4.11}$$

$$\tilde{\gamma}_j^n = \sum_{l=1}^{j} \big[(\psi_l^n - \kappa_{l-1}^n) + (\kappa_l^n - \phi_l^n)\big], \qquad 1 \le j \le \nu^n, \tilde{\gamma}_0^n = 0,$$
$$\tilde{\beta}_j^n = \tilde{\gamma}_{j-1}^n + \big(\psi_j^n - \kappa_{j-1}^n\big), \qquad 1 \le j \le \nu^n, \tilde{\beta}_0^n = 0. \tag{4.12}$$

The exceptional interarrival times for this queue are $v_1^n, v_2^n, \dots$. That is, these are the interarrival times of the first arrivals in busy periods.

Equalities (4.11) and (4.12) and the assumption $f \ge 0$ show that

$$\int_0^{\tilde{\vartheta}^n} f\big(\tilde{X}_1^n(t)\big)\,dt \le \int_{\tau_k^n(\varepsilon, i)\wedge T}^{\zeta_k^n(\varepsilon, i)\wedge T} f(X_1^n(t))\,dt, \tag{4.13}$$

where $\tilde{\vartheta}^n = \tilde{\gamma}_{\nu^n}^n$, that is, $\tilde{X}_1^n$ represents a lower-bound process.

Now we construct an upper-bound process $\hat{X}_1^n$. Whereas we truncated the original process X_1^n at the level $a_i(\varepsilon) - \varepsilon$ on $[\tau_k^n(\varepsilon, i) \wedge T, \zeta_k^n(\varepsilon, i) \wedge T]$ to obtain the lower-bound process, here we will extend X_1^n on that interval to level $a_i(\varepsilon) + \varepsilon$ to obtain an upper-bound process (see Figure 4). Introduce independent replicas $\{\hat{\xi}_{j,l}^{n,m}, l \geq 1\}$, $j \geq 1$, $m = 1, 2$, of the interarrival-time sequence and independent replicas $\{\hat{\eta}_{j,l}^n, l \geq 1\}$, $j \geq 1$, of the service-time sequence.

Let

$$\varphi_j^n = \inf\big(t > \kappa_{j-1}^n : X_1^n(t) > a_i(\varepsilon) + \varepsilon\big) \wedge \theta_j^n, \qquad j \geq 1,$$

$$\vartheta_j^n = \inf\left(t > \theta_j^n : X_1^n(t) \leq a_i(\varepsilon) + \varepsilon + \frac{1}{\sqrt{n}}\right), \qquad j \geq 1.$$

Note that $Q_1^n(n\varphi_j^n) = Q_1^n(n\vartheta_j^n)$ and if $X_1^n(\theta_j^n) \leq a_i(\varepsilon) + \varepsilon$, then $\varphi_j^n = \vartheta_j^n = \theta_j^n$. Let $\hat{\chi}_j^n$, $j \geq 1$, index the arrival in the original queue occurring at or just after $n\varphi_j^n$ (recall that the numbering starts from the arrival after $n\kappa_0^n$) and let $\bar{v}_j^n$, $j \geq 1$, denote the time between $n\varphi_j^n$ and the $\hat{\chi}_j^n$th arrival. By definition, $\bar{v}_j^n \leq \tilde{\xi}_{\hat{\chi}_j^n}^n$. Also $\bar{v}_j^n = 0$ if $\varphi_j^n < \theta_j^n$.

Construct as follows a threshold queue with the threshold $h^n = \lfloor \sqrt{n}\,(a_i(\varepsilon) + \varepsilon) \rfloor$. In the first cycle the interarrival times in the accumulation period are taken from the sequence $\{\tilde{\xi}_1^n, \tilde{\xi}_2^n, \dots, \tilde{\xi}_{\hat{\chi}_1^n}^n, \hat{\xi}_{1,1}^{n,1}, \hat{\xi}_{1,2}^{n,1}, \dots\}$ [note that if $Q_1^n(n\varphi_1^n) = h^n + 1$, then $\hat{\xi}_{1,1}^{n,1}, \hat{\xi}_{1,2}^{n,1}, \dots$ are not used]. Denoting the threshold queue length at t by $\hat{Q}_1^n(t)$, define

$$\hat{\beta}_1^n = \inf\big(t > 0 : \hat{Q}_1^n(nt) > h^n\big).$$

Then $n\hat{\beta}_1^n$ ends the accumulation period. If $Q_1^n(n\varphi_1^n) \leq h^n$, which happens if

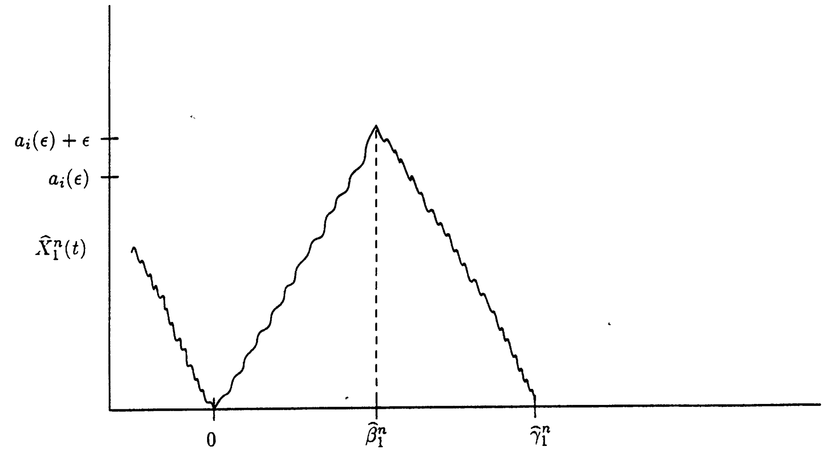

FIG. 4. *The upper-bound process.*

$Q_1^n(n\theta_1^n) \le h^n$, then after $n\hat{\beta}_1^n$ the service times are $\hat{\eta}_{1,1}^n, \hat{\eta}_{1,2}^n, \dots,$ and the initial interarrival times are taken from $\{\hat{\xi}_{1,l}^{n,2},\ l \ge 1\}$ until time $n\breve{\beta}_1^n$, where

$$\breve{\beta}_1^n = \inf\left(t > \hat{\beta}_1^n : \hat{Q}_1^n(nt) = Q_1^n(n\vartheta_1^n)\right).$$

We then take $\hat{\xi}_{1,\bar{\chi}_1^n}^{n,2}$ to be the last random variable in the sequence $\{\hat{\xi}_{1,1}^{n,2}, \hat{\xi}_{1,2}^{n,2}, \dots\}$ that was actually realized as an interarrival time in $[n\hat{\beta}_1^n, n\breve{\beta}_1^n]$. In the case that $Q_1^n(n\varphi_1^n) = h^n + 1$, which happens if $Q_1^n(n\theta_1^n) > h^n$, we define $\breve{\beta}_1^n = \hat{\beta}_1^n$ and set $\bar{\chi}_1^n = \hat{\xi}_{1,\bar{\chi}_1^n}^{n,2} = 0$. In both cases, the first arrival after $n\breve{\beta}_1^n$ is made to occur at time $n\breve{\beta}_1^n + \bar{v}_1^n$, so that its interarrival time $\hat{v}_1^n$ always satisfies $\hat{v}_1^n \le \hat{\xi}_{1,\bar{\chi}_1^n+1}^{n,2} + \bar{v}_1^n \le \hat{\xi}_{1,\bar{\chi}_1^n+1}^{n,2} + \tilde{\xi}_{\hat{\chi}_1^n}^n$. The subsequent interarrival times are $\tilde{\xi}_{\hat{\chi}_1^n+1}^n, \dots, \tilde{\xi}_{\hat{\chi}_2^n}^n$, and the service times after $n\breve{\beta}_1^n$ are the same as for Q_1^n after $n\vartheta_1^n$. The arrival terminating the interarrival time $\tilde{\xi}_{\hat{\chi}_2^n}^n$ corresponds to the arrival in the original queue occurring at or after $n\varphi_2^n$. After that arrival, the interarrival times are again taken to be $\hat{\xi}_{2,1}^{n,1}, \hat{\xi}_{2,2}^{n,1}, \dots$ until the threshold has been exceeded [these times are not used if $Q_1^n(n\varphi_2^n) = h^n + 1$]. After this has happened at $n\hat{\beta}_2^n$, where

$$\hat{\beta}_2^n = \inf\left(t > \hat{\beta}_1^n : \hat{Q}_1^n(nt) > h^n\right),$$

and until $n\breve{\beta}_2^n$, where

$$\breve{\beta}_2^n = \inf\left(t > \hat{\beta}_2^n : \hat{Q}_1^n(nt) = Q_1^n(n\vartheta_2^n)\right),$$

the service times are $\hat{\eta}_{2,1}^n, \hat{\eta}_{2,2}^n, \dots$ and the interarrival times are $\hat{\xi}_{2,1}^{n,2}, \hat{\xi}_{2,2}^{n,2}, \dots$ [as above, these are not used if $Q_1^n(n\varphi_2^n) = h^n + 1$ and hence $\hat{\beta}_2^n = \breve{\beta}_2^n$]. After $n\breve{\beta}_2^n$, the next arrival occurs at time $n\breve{\beta}_2^n + \bar{v}_2^n$ [in both cases, $Q_1^n(n\varphi_2^n) = h^n + 1$ and $Q_1^n(n\varphi_2^n) \le h^n$], so that its interarrival time satisfies $\hat{v}_2^n \le \hat{\xi}_{2,\bar{\chi}_2^n+1}^{n,2} + \bar{v}_2^n \le \hat{\xi}_{2,\bar{\chi}_2^n+1}^{n,2} + \tilde{\xi}_{\hat{\chi}_2^n}^n$, where $\hat{\xi}_{2,\bar{\chi}_2^n}^{n,2}$ is the last random variable from $\{\hat{\xi}_{2,1}^{n,2}, \hat{\xi}_{2,2}^{n,2}, \dots\}$ that is realized as an interarrival time in $[n\hat{\beta}_2^n, n\breve{\beta}_2^n]$ [again $\bar{\chi}_2^n = \hat{\xi}_{2,\bar{\chi}_2^n}^{n,2} = 0$ if $Q_1^n(n\varphi_2^n) = h^n + 1$ and hence $\hat{\beta}_2^n = \breve{\beta}_2^n$]. The subsequent interarrival times are $\tilde{\xi}_{\hat{\chi}_2^n+1}^n, \dots, \tilde{\xi}_{\hat{\chi}_3^n}^n$, and the service times replicate those after $n\vartheta_2^n$. After the last of the above arrivals the cycle resumes.

That this is indeed a threshold queue with generic interarrival and service times distributed as in the original queue follows by the next lemma; the proof is routine and left to the reader.

LEMMA 4.2. *Let $\{\xi_i^1,\ i \ge 1\}$ and $\{\xi_i^2,\ i \ge 1\}$ be identically distributed i.i.d. sequences. Let ξ_i^1, $i = 1, 2, \dots,$ be measurable with respect to a σ-field $\mathscr{F}_i$, the latter being independent of $\xi_{i+1}^1, \xi_{i+2}^1, \dots,$ and of $\{\xi_i^2,\ i \ge 1\}$. If $\chi = 0, 1, \dots$ is a stopping time with respect to the flow $(\mathscr{F}_i,\ i \ge 0)$, then the sequence*

$\{\xi_1^1, \dots, \xi_\chi^1, \xi_1^2, \xi_2^2, \dots\}$ *has the same distribution as* $\{\xi_i^1,\ i \geq 1\}$, *where by definition* $\{\xi_1^1, \dots, \xi_\chi^1, \xi_1^2, \xi_2^2, \dots\} = \{\xi_1^2, \xi_2^2, \dots\}$, *if* $\chi = 0$.

It is readily seen, as an application of the lemma, that $\{\tilde{\xi}_2^n, \dots, \tilde{\xi}_{\hat{\chi}_1^n}^n, \hat{\xi}_{1,1}^{n,1}, \hat{\xi}_{1,2}^{n,1}, \dots\}$ and $\{\hat{\xi}_{1,1}^{n,2}, \dots, \hat{\xi}_{1,\bar{\chi}_1^n}^{n,2}, \tilde{\xi}_{\hat{\chi}_1^n+1}^n, \tilde{\xi}_{\hat{\chi}_1^n+2}^n, \dots\}$ are i.i.d. sequences, and that the service times in each busy period form an i.i.d. sequence. This proves that the interarrival and service times follow the same laws as in the original queue, except for the interarrival periods $\tilde{\xi}_1^n, \hat{v}_1^n, \hat{v}_2^n, \dots$. Note, however, that the two interarrival sequences above are generally dependent. Still, this will not prevent us from applying Theorem 3.1 in view of the remark after its proof.

Define $\hat{X}_1^n(t) = \hat{Q}_1^n(nt)/\sqrt{n}$ to be the normalized and scaled queue length in the above threshold queue, and let $\hat{\beta}_i^n$ and $\hat{\gamma}_i^n$ be defined for this queue as in (3.4) (which agrees with the earlier definition of $\hat{\beta}_1^n$ and $\hat{\beta}_2^n$). By construction,

$$
\begin{aligned}
&\int_{\tau_k^n(\varepsilon, i)\wedge T}^{\zeta_k^n(\varepsilon, i)\wedge T} f(X_1^n(t))\,dt \\
&\qquad \leq \int_{\tau_k^n(\varepsilon, i)\wedge T}^{\kappa_0^n\wedge T} f(X_1^n(t))\,dt \\
&\qquad\qquad + \int_{\kappa_{\nu^n}^n\wedge T}^{\zeta_k^n(\varepsilon, i)\wedge T} f(X_1^n(t))\,dt + \int_0^{\hat{\vartheta}^n} f\big(\hat{X}_1^n(t)\big)\,dt,
\end{aligned}
\tag{4.14}
$$

where $\hat{\vartheta}^n = \hat{\gamma}_{\nu^n}^n$.

We now check that $\tilde{X}_1^n$ and $\hat{X}_1^n$ satisfy the conditions of Theorem 3.1. We need focus only on the parts related to exceptional interarrival times and the times of the first arrivals. Define

$$
\tilde{\zeta}_j^n(r) = \max_{1 \leq l \leq \lfloor \sqrt{n}\, r \rfloor} \tilde{\xi}_{\tilde{\chi}_j^{n,1}+l}^n, \qquad j \geq 1,\ r > 0.
$$

Noting that $\{\tilde{\xi}_{\tilde{\chi}_j^{n,1}+l}^n,\ l \geq 1\}$ is distributed as $\{\tilde{\xi}_l^n,\ l \geq 1\}$ and that $v_j^n \leq \tilde{\xi}_{\tilde{\chi}_j^{n,2}}^n$, one can prove in analogy with Lemma 3.1 that $\{v_j^n, \tilde{\zeta}_j^n(r),\ j \geq 1\}$ satisfies the conditions of Theorem 3.1. Similarly, $\tilde{\xi}_1^n/\sqrt{n} \to_P 0$. Thus, the conditions of Theorem 3.1 hold for the lower-bound process.

For the exceptional interarrival times $\{\hat{v}_j^n,\ j \geq 1\}$ in the upper-bound process, the argument uses the random variables

$$
\hat{\zeta}_j^n(r) = \max_{1 \leq l \leq \lfloor \sqrt{n}\, r \rfloor} \tilde{\xi}_{\breve{\chi}_j^n+l} + \max_{1 \leq l \leq \lfloor \sqrt{n}\, r \rfloor} \hat{\xi}_{j,l}^{n,2},
$$

where $\breve{\chi}_j^n$ indexes the first arrival in the original queue after κ_{j-1}^n, and the inequality $\hat{v}_j^n \leq \tilde{\xi}_{\hat{\chi}_j^n} + \hat{\xi}_{1,\bar{\chi}_j^n+1}^{n,2}$. Again a formal proof is worked out as in Lemma 3.1. The first interarrival time is again $\tilde{\xi}_1^n$. We conclude that Theorem 3.1 holds for $\tilde{X}_1^n$ and $\hat{X}_1^n$.

Next, by (4.13) and (4.14) we have the bounds

$$
\begin{aligned}
&\int_0^{\tilde{\vartheta}^n} f\big(\tilde{X}_1^n(t)\big)\,dt\\
&\quad\le \int_{\tau_k^n(\varepsilon,i)\wedge T}^{\zeta_k^n(\varepsilon,i)\wedge T} f\big(X_1^n(t)\big)\,dt \qquad (4.15)\\
&\quad\le \int_0^{\hat{\vartheta}^n} f\big(\hat{X}_1^n(t)\big)\,dt + \big[\kappa_0^n \wedge T - \tau_k^n(\varepsilon,i)\wedge T\big]\|f\|\\
&\qquad + \big[\zeta_k^n(\varepsilon,i)\wedge T - \kappa_{\nu^n}^n \wedge T\big]\|f\|.
\end{aligned}
$$

Define

$$
\begin{aligned}
\tilde{\nu}^n &= \min\big(j\colon \tilde{\gamma}_{j+1}^n > \zeta_k^n(\varepsilon,i)\wedge T - \tau_k^n(\varepsilon,i)\wedge T\big),\\
\hat{\nu}^n &= \min\big(j\colon \hat{\gamma}_{j+1}^n > \zeta_k^n(\varepsilon,i)\wedge T - \tau_k^n(\varepsilon,i)\wedge T\big),
\end{aligned}
$$

and let $U_k^n(\varepsilon,i)$ and $V_k^n(\varepsilon,i)$ denote, respectively, the lower bound in (4.15) with $\tilde{\vartheta}^n$ changed to $\tilde{\omega}^n = \tilde{\gamma}_{\tilde{\nu}^n}^n$ and the upper bound in (4.15) with $\hat{\vartheta}^n$ changed to $\hat{\omega}^n = \hat{\gamma}_{\hat{\nu}^n}^n$. Since obviously $\hat{\nu}^n \le \nu^n \le \tilde{\nu}^n$, we have that

$$U_k^n(\varepsilon,i) \le \int_{\tau_k^n(\varepsilon,i)\wedge T}^{\zeta_k^n(\varepsilon,i)\wedge T} f\big(X_1^n(t)\big)\,dt \le V_k^n(\varepsilon,i). \tag{4.16}$$

We now show that

$$U_k^n(\varepsilon,i) \to_d U_k(\varepsilon,i), \qquad V_k^n(\varepsilon,i) \to_d V_k(\varepsilon,i), \qquad k \ge 1,\ 0 \le i \le N, \tag{4.17}$$

where

$$
\begin{aligned}
U_k(\varepsilon,i) &= \frac{a_i(\varepsilon)-\varepsilon}{a_i(\varepsilon)+\varepsilon}\big(\zeta_k(\varepsilon,i)\wedge T - \tau_k(\varepsilon,i)\wedge T\big)\\
&\quad\times\int_0^1 f\big(u(a_i(\varepsilon)-\varepsilon)\big)\,du,\\
V_k(\varepsilon,i) &= \frac{a_i(\varepsilon)+\varepsilon}{a_i(\varepsilon)-\varepsilon}\big(\zeta_k(\varepsilon,i)\wedge T - \tau_k(\varepsilon,i)\wedge T\big)\\
&\quad\times\int_0^1 f\big(u(a_i(\varepsilon)+\varepsilon)\big)\,du.
\end{aligned}
\tag{4.18}
$$

Let $\tilde{\nu}^n(t) = \min(j \ge 0\colon \tilde{\gamma}_{j+1}^n > t)$ and $\hat{\nu}^n(t) = \min(j \ge 0\colon \hat{\gamma}_{j+1}^n > t)$. In the course of proving Theorem 3.1 we established (3.5). Since $\tilde{X}_1^n$ and $\hat{X}_1^n$ meet the conditions of Theorem 3.1, we can write for these processes, in analogy with (3.5),

$$
\begin{aligned}
\tilde{\gamma}_{\lfloor\sqrt{n}t\rfloor}^n &\to_P t\big(a_i(\varepsilon)-\varepsilon\big)\left(\frac{1}{\lambda_1}+\frac{1}{\mu-\lambda_1}\right),\\
\hat{\gamma}_{\lfloor\sqrt{n}t\rfloor}^n &\to_P t\big(a_i(\varepsilon)+\varepsilon\big)\left(\frac{1}{\lambda_1}+\frac{1}{\mu-\lambda_1}\right).
\end{aligned}
$$

Then Lemma 2.1 yields

$$\frac{\tilde{\nu}^n(t)}{\sqrt{n}} \to_P \frac{t}{a_i(\varepsilon) - \varepsilon}\left(\frac{1}{\lambda_1} + \frac{1}{\mu - \lambda_1}\right)^{-1},$$

$$\frac{\hat{\nu}^n(t)}{\sqrt{n}} \to_P \frac{t}{a_i(\varepsilon) + \varepsilon}\left(\frac{1}{\lambda_1} + \frac{1}{\mu - \lambda_1}\right)^{-1},$$

whence (by Lemma 2.2, for example),

$$\tilde{\gamma}^n_{\tilde{\nu}^n(t)} \to_P \frac{a_i(\varepsilon) - \varepsilon}{a_i(\varepsilon) + \varepsilon} t, \qquad \hat{\gamma}^n_{\hat{\nu}^n(t)} \to_P \frac{a_i(\varepsilon) + \varepsilon}{a_i(\varepsilon) - \varepsilon} t.$$

Then by Theorem 3.1,

$$\begin{aligned} &\int_0^{\tilde{\gamma}^n_{\tilde{\nu}^n(t)}} f(\tilde{X}_1^n(s))\,ds \to_P \frac{a_i(\varepsilon) - \varepsilon}{a_i(\varepsilon) + \varepsilon} t \int_0^1 f(u(a_i(\varepsilon) - \varepsilon))\,du, \\ &\int_0^{\hat{\gamma}^n_{\hat{\nu}^n(t)}} f(\hat{X}_1^n(s))\,ds \to_P \frac{a_i(\varepsilon) + \varepsilon}{a_i(\varepsilon) - \varepsilon} t \int_0^1 f(u(a_i(\varepsilon) + \varepsilon))\,du. \end{aligned} \tag{4.19}$$

Since $\tilde{\nu}^n = \tilde{\nu}^n(\zeta_k^n(\varepsilon, i) \wedge T - \tau_k^n(\varepsilon, i) \wedge T)$ and $\hat{\nu}^n = \hat{\nu}^n(\zeta_k^n(\varepsilon, i) \wedge T - \tau_k^n(\varepsilon, i) \wedge T)$, Lemmas 2.2 and 4.1 show that (4.19) implies

$$\begin{aligned} &\int_0^{\tilde{\omega}^n} f(\tilde{X}_1^n(s))\,ds \\ &\quad \to_d \frac{a_i(\varepsilon) - \varepsilon}{a_i(\varepsilon) + \varepsilon}(\zeta_k(\varepsilon, i) \wedge T - \tau_k(\varepsilon, i) \wedge T) \int_0^1 f(u(a_i(\varepsilon) - \varepsilon))\,du, \\ &\int_0^{\hat{\omega}^n} f(\hat{X}_1^n(s))\,ds \\ &\quad \to_d \frac{a_i(\varepsilon) + \varepsilon}{a_i(\varepsilon) - \varepsilon}(\zeta_k(\varepsilon, i) \wedge T - \tau_k(\varepsilon, i) \wedge T) \int_0^1 f(u(a_i(\varepsilon) + \varepsilon))\,du. \end{aligned}$$

Since $|\kappa_0^n \wedge T - \tau_k^n(\varepsilon, i) \wedge T| \to_P 0$ and $|\zeta_k^n(\varepsilon, i) \wedge T - \kappa_{\nu^n}^n \wedge T| \to_P 0$ obviously hold, (4.17) is proved. Moreover, the same argument shows that

$$\begin{aligned} &\left(X^n, (U_k^n(\varepsilon, i))_{k \ge 1, 0 \le i \le N}\right) \to_d \left(X, (U_k(\varepsilon, i))_{k \ge 1, 0 \le i \le N}\right), \\ &\left(X^n, (V_k^n(\varepsilon, i))_{k \ge 1, 0 \le i \le N}\right) \to_d \left(X, (V^k(\varepsilon, i))_{k \ge 1, 0 \le i \le N}\right). \end{aligned} \tag{4.20}$$

Next, defining

$$U^n(\varepsilon) = \sum_{k=1}^{\infty} \sum_{i=0}^{N} U_k^n(\varepsilon, i), \qquad V^n(\varepsilon) = \sum_{k=1}^{\infty} \sum_{i=0}^{N} V_k^n(\varepsilon, i), \tag{4.21}$$

we need to prove that

$$(X^n, U^n(\varepsilon)) \to_d (X, U(\varepsilon)), \qquad (X^n, V^n(\varepsilon)) \to_d (X, V(\varepsilon)), \tag{4.22}$$

where

$$U(\varepsilon) = \sum_{k=1}^{\infty} \sum_{i=0}^{N} U_k(\varepsilon, i), \qquad V(\varepsilon) = \sum_{k=1}^{\infty} \sum_{i=0}^{N} V_k(\varepsilon, i). \tag{4.23}$$

We prove the first convergence result in (4.22); the proof of the second uses the same reasoning.

Since $U_k^n(\varepsilon, i) = 0$ if $\tau_k^n(\varepsilon, i) \geq T$, in view of (4.21) we have by Lemma 4.1, for $\eta > 0$,

$$\begin{aligned} &\limsup_{M\to\infty} \limsup_{n\to\infty} P\left(\left| \sum_{k=1}^{M} \sum_{i=0}^{N} U_k^n(\varepsilon, i) - U^n(\varepsilon) \right| > \eta \right) \\ &\leq \limsup_{M\to\infty} \limsup_{n\to\infty} P\left(\min_{0\leq i\leq N} \zeta_M^n(\varepsilon, i) \wedge (T+1) < T \right) \qquad (4.24) \\ &\leq \limsup_{M\to\infty} P\left(\min_{0\leq i\leq N} \zeta_M(\varepsilon, i) \wedge (T+1) \leq T \right) = 0. \end{aligned}$$

Analogously,

$$\sum_{k=1}^{M} \sum_{i=0}^{N} U_k(\varepsilon, i) \to_P U(\varepsilon), \qquad M \to \infty. \tag{4.25}$$

Next, by (4.20), we have

$$\left(X^n, \sum_{k=1}^{M} \sum_{i=0}^{N} U_k^n(\varepsilon, i) \right) \to_d \left(X, \sum_{k=1}^{M} \sum_{i=0}^{N} U_k(\varepsilon, i) \right). \tag{4.26}$$

The convergence $(X^n, U^n(\varepsilon)) \to_d (X, U(\varepsilon))$ then follows from (4.23)–(4.26) and Theorem 4.2 in Billingsley (1968).

Now by (4.7) and (4.8),

$$\begin{aligned} &\left| \int_0^T f(X_1^n(t)) \cdot 1(\delta \leq X^n(t) \leq K)\, dt \right. \\ &\left. - \sum_{k=1}^{\infty} \sum_{i=0}^{N} \int_0^T f(X_1^n(t)) \cdot 1(t \in [\tau_k^n(\varepsilon, i), \zeta_k^n(\varepsilon, i)))\, dt \right| \\ &\leq \|f\| \int_0^T [1(\delta - \varepsilon \leq X^n(t) \leq \delta) + 1(K \leq X^n(t) \leq K + \varepsilon)]\, dt, \end{aligned}$$

so by (4.16), we obtain from (4.21),

$$\begin{aligned} U^n(\varepsilon) - \|f\| \int_0^T &[1(\delta - \varepsilon \leq X^n(t) \leq \delta) \\ &+ 1(K \leq X^n(t) \leq K + \varepsilon)]\, dt \\ &\leq \int_0^T f(X_1^n(t)) \cdot 1(\delta \leq X^n(t) \leq K)\, dt \qquad (4.27) \\ &\leq V^n(\varepsilon) + \|f\| \int_0^T [1(\delta - \varepsilon \leq X^n(t) \leq \delta) \\ &\qquad + 1(K \leq X^n(t) \leq K + \varepsilon)]\, dt. \end{aligned}$$

Therefore, if we prove that, as $\varepsilon \to 0$,

$$
\begin{aligned}
&U(\varepsilon) \to_d \int_0^T \left(\int_0^1 f(uX(t))\, du \right) 1(\delta \le X(t) \le K)\, dt, \\
&V(\varepsilon) \to_d \int_0^T \left(\int_0^1 f(uX(t))\, du \right) 1(\delta \le X(t) \le K)\, dt,
\end{aligned} \tag{4.28}
$$

then by applying Lemma 2.3 to (4.27) and taking into account (4.22) and the fact that $\int_0^T 1(X(t) = a)\, dt = 0$ P-a.s., $a > 0$, we will then obtain (4.1) and hence the assertion of Theorem 2.1. As before, we prove only the first of the results in (4.28); the proof of the second is similar.

In fact, we prove convergence with probability 1. Since $a_i(\varepsilon) > \delta$, we have from (4.18) and (4.23),

$$
\begin{aligned}
&\left| U(\varepsilon) - \sum_{k=1}^{\infty} \sum_{i=0}^{N} [\zeta_k(\varepsilon, i) \wedge T - \tau_k(\varepsilon, i) \wedge T] \int_0^1 f(u(a_i(\varepsilon) - \varepsilon))\, du \right| \\
&\quad \le \frac{2\varepsilon}{\delta} \|f\| T.
\end{aligned}
$$

This tends to 0 as $\varepsilon \to 0$, so we prove that, with probability 1,

$$
\begin{aligned}
&\lim_{\varepsilon \to 0} \sum_{k=1}^{\infty} \sum_{i=0}^{N} [\zeta_k(\varepsilon, i) \wedge T - \tau_k(\varepsilon, i) \wedge T] \\
&\qquad \times \int_0^1 f(u(a_i(\varepsilon) - \varepsilon))\, du \\
&\qquad = \int_0^T \left(\int_0^1 f(uX(t))\, du \right) 1(\delta \le X(t) \le K)\, dt.
\end{aligned} \tag{4.29}
$$

We can write

$$
\begin{aligned}
&\sum_{k=1}^{\infty} \sum_{i=0}^{N} [\zeta_k(\varepsilon, i) \wedge T - \tau_k(\varepsilon, i) \wedge T] \int_0^1 f(u(a_i(\varepsilon) - \varepsilon))\, du \\
&\quad = \sum_{k=1}^{\infty} \sum_{i=0}^{N} \int_0^T \left(\int_0^1 f(u(a_i(\varepsilon) - \varepsilon))\, du \right) \\
&\qquad\qquad \times 1(\tau_k(\varepsilon, i) \le t < \zeta_k(\varepsilon, i))\, dt \\
&\quad \equiv C_\varepsilon.
\end{aligned} \tag{4.30}
$$

Note that if $x, y > \delta/2$, $|x - y| < 2\varepsilon$, then

$$
\begin{aligned}
\left| \int_0^1 f(ux)\, du - \int_0^1 f(uy)\, du \right| &= \left| \frac{1}{x} \int_0^x f(u)\, du - \frac{1}{y} \int_0^y f(u)\, du \right| \\
&\le \left| \frac{1}{x} - \frac{1}{y} \right| \left| \int_0^x f(u)\, du \right| + \frac{1}{y} \left| \int_x^y f(u)\, du \right| \\
&\le \frac{8\varepsilon}{\delta} \|f\|.
\end{aligned}
$$

Since $X(t) \in [a_i(\varepsilon) - \varepsilon,\ a_i(\varepsilon) + \varepsilon]$ if $t \in [\tau_k(\varepsilon, i),\ \zeta_k(\varepsilon, i))$, we then have

$$\begin{aligned}&\Big|\int_0^1 f(u(a_i(\varepsilon) - \varepsilon))\,du \cdot 1(t \in [\tau_k(\varepsilon,i), \zeta_k(\varepsilon,i)))\\ &\qquad - \int_0^1 f(uX(t))\,du \cdot 1(t \in [\tau_k(\varepsilon,i), \zeta_k(\varepsilon,i)))\Big|\\ &\qquad \le \frac{8\varepsilon}{\delta}\|f\| \cdot 1(t \in [\tau_k(\varepsilon,i), \zeta_k(\varepsilon,i))),\end{aligned}$$

so by (4.30),

$$\begin{aligned}&\left|C_\varepsilon - \sum_{k=1}^{\infty}\sum_{i=0}^{N}\int_0^T\left(\int_0^1 f(uX(t))\,du\right) \cdot 1(t \in [\tau_k(\varepsilon,i), \zeta_k(\varepsilon,i)))\,dt\right|\\ &\qquad \le \frac{8\varepsilon}{\delta}\|f\|T,\end{aligned}$$

whence, by (4.8) expressed in terms of the limit process $(X(t),\ t \ge 0)$,

$$\begin{aligned}&\left|C_\varepsilon - \int_0^T\left(\int_0^1 f(uX(t))\,du\right) \cdot 1(\delta \le X(t) \le K)\,dt\right|\\ &\qquad \le \|f\|\int_0^T[1(\delta \le X(t) \le \delta - \varepsilon) + 1(K \le X(t) \le K + \varepsilon)]\,dt + \frac{8\varepsilon}{\delta}\|f\|T.\end{aligned}$$

Since the right-hand side of this inequality tends to 0 with probability 1 as $\varepsilon \to 0$, we have proved (4.29). This completes the proof of Theorem 2.1 for bounded nonnegative, continuous f. The claim for bounded, continuous f follows since $f(x) - \inf_y f(y)$ is nonnegative.

Finally, the case of unbounded, continuous f is treated via a localization argument. Define, for $A > 0$,

$$\sigma_A^n = \inf(t > 0\colon X_1^n(t) > A).$$

Since

$$\sup_{s \le t} X_1^n(s) \le \sup_{s \le t} X^n(s)$$

and

$$\sup_{s \le t} X^n(s) \to_d \sup_{s \le t} Y(s),$$

we have

$$\lim_{A \to \infty}\limsup_{n \to \infty} P(\sigma_A^n < T) = 0. \tag{4.31}$$

Let $f_A(x) = f(x \wedge A)$, $x \ge 0$. Since f_A is bounded as a consequence of the continuity of f,

$$\int_0^T f_A(X_1^n(t))\,dt \to_d \int_0^T\left(\int_0^1 f_A(uX(t))\,du\right)dt$$

by the case already proved. Further,

$$\left|\int_0^T f_A(X_1^n(t))\,dt - \int_0^T f(X_1^n(t))\,dt\right| \le \int_0^T |f(X_1^n(t))|\cdot 1(|X_1^n(t)| > A)\,dt$$
$$\le \int_{T\wedge\sigma_A^n}^T |f(X_1^n(t))|dt,$$

and, by (4.31),

$$\lim_{A\to\infty}\limsup_{n\to\infty} P\left(\left|\int_0^T f_A(X_1^n(t))\,dt - \int_0^T f(X_1^n(t))\,dt\right| > 0\right) = 0.$$

Noting also that, as $A \to \infty$,

$$\int_0^T\left(\int_0^1 f_A(uX(t))\,du\right)dt \to \int_0^T\left(\int_0^1 f(uX(t))\,du\right)dt, \qquad P\text{-a.s.},$$

we conclude by Theorem 4.2 in Billingsley (1968) that

$$\int_0^T f(X_1^n(t))\,dt \to_d \int_0^T\left(\int_0^1 f(uX(t))\,du\right)dt. \qquad \square$$

REMARKS. 1. Arguments similar to those used in the proof of Theorem 2.1 also show that finite-dimensional convergence holds. That is,

$$\begin{aligned}&\left(\int_0^{t_1} f(X_l^n(s))\,ds,\dots,\int_0^{t_k} f(X_l^n(s))\,ds\right)\\ &\quad\to_d \left(\int_0^{t_1}\left(\int_0^1 f(uX(s))\,du\right)ds,\dots,\right.\\ &\qquad\qquad \left.\int_0^{t_k}\left(\int_0^1 f(uX(s))\,du\right)ds\right).\end{aligned}\tag{4.32}$$

Therefore, since the sequence $Y_l^n = (\int_0^t f(X_l^n(s))\,ds,\ t \ge 0)$ is seen to be tight, we have the functional convergence in $C[0,\infty)$, $Y_l^n \to_d Y$, where $Y = (\int_0^t (\int_0^1 f(uX(s))\,du)\,ds,\ t \ge 0)$.

2. In the course of proving Theorem 2.1, we showed that in fact, for all $0 < \delta < K$ and bounded, nonnegative and continuous f,

$$\int_0^T f(X_l^n(t))\cdot 1(\delta \le X^n(t) \le K)\,dt$$
$$\to_d \int_0^T\left(\int_0^1 f(uX(t))\,du\right)\cdot 1(\delta \le X(t) \le K)\,dt.$$

From this it easily follows that, for any function $g(x,y)$ continuous in both variables, we have

$$\int_0^T g(X^n(t), X_l^n(t))\,dt \to_d \int_0^T\left(\int_0^1 g(X(t), uX(t))\,du\right)dt.$$

[To verify this result, introduce $b > 0$ and $0 < a_1 < a_2$ and use

$$\lim_{N\to\infty} \sup_{y\le b} \sup_{a_1\le x\le a_2} \left| g(x,y) - \sum_{i=1}^{N} g(a_1 + \varepsilon i, y) \times 1\big(x \in [a_1 + \varepsilon(i-1), a_1 + \varepsilon i)\big) \right| = 0,$$

where $\varepsilon = (a_2 - a_1)/N$.] Since $X_2^n(t) = X^n(t) - X_1^n(t)$, we then obtain, taking (4.32) into account as well, that

$$\int g(X_1^n(t), X_2^n(t))\,dt \to_d \int\left(\int_0^1 g(uX(t),(1-u)X(t))\,du\right)dt,$$

the convergence being functional convergence in $C[0,\infty)$.

We can now formalize the waiting-time averaging arguments illustrated in the Introduction. Let $W_l^n(t)$ denote the virtual waiting time, that is, the waiting time that a customer arriving at time t would have, and define $Z_l^n(t) = (1/\sqrt{n})W_l^n(nt)$, $l = 1,2$, $n \ge 1$, $t \ge 0$. From Theorem 2.1, in analogy with Theorem 3.2, we have the following averaging principle for virtual waiting times.

THEOREM 4.1. *Under the conditions of Theorem* 2.1,

$$\int_0^T f(Z_l^n(t))\,dt \to_d \int_0^T\int_0^1 f(uX(t)/\lambda_l)\,du\,dt, \qquad l = 1,2,$$

as $n \to \infty$.

With $V(t) = X(t)/\mu$ and U a uniform random variable on $[0,1]$ independent of $V(t)$, the integral over u in the limit of Theorem 4.1 can be written as the conditional expectation

$$E\big[f(UV(t)/\rho_l)|V(t)\big] = \int_0^1 f(uV(t)/\rho_l)\,du. \tag{4.33}$$

If we take the symmetric case $\rho_1 = \rho_2 = 1/2$ and set $f(x) = x$, then (4.33) agrees with (1.1). Theorem 4.1 and (4.33) also show that, if f is bounded,

$$\int_0^T Ef(Z_l^n(t))\,dt \to \int_0^T Ef(UV(t)/\rho_l)\,dt, \qquad l = 1,2. \tag{4.34}$$

As a final comment, we note that, as is common in heavy-traffic limit theorems, if the actual waiting time prccess is defined appropriately, then it converges together with the virtual waiting-time process. In particular, if we let $\tilde{W}_l(t)$ denote the actual waiting (or sojourn) time of the first customer to arrive at queue l after time t, then for any T, $0 < T < \infty$,

$$\sup_{0\le t\le T} \frac{1}{\sqrt{n}}\left|W_l^n(nt) - \tilde{W}_l^n(nt)\right| \to_P 0, \qquad n \to \infty.$$

5. Extensions. To this point, it has been convenient to have the same service-time distribution at both queues, but this assumption is not essential. The results are readily adapted to a fully asymmetric system with queue-dependent service times; however, the "conserved" process is the total unfinished work rather than the total queue length. To be more specific, we need further notation. (Because of the heavy demand for notation in Sections 3 and 4, the remainder of the paper reuses some of the symbols for other purposes. The conflicts should cause no difficulty, since there will be no further need to refer to the proof details in Sections 3 and 4.)

Let $(\sigma_{al}^n)^2$, $l = 1,2$, denote the arrival-time variance at queue l, with $(\sigma_{al}^n)^2 \to \sigma_{al}^2$ $(n \to \infty)$, and introduce $\mu_l^n, (\sigma_{sl}^n)^2$, $l = 1,2$, as the respective service rate and service-time variance at queue l, with $\lim_{n\to\infty} \mu_l^n = \mu_l$, $\lim_{n\to\infty} (\sigma_{sl}^n)^2 = \sigma_{sl}^2$ and $\rho_l^n = \lambda_l^n/\mu_l^n$. Let $V_l^n(t) = (1/\sqrt{n})L_l^n(nt)$, $l = 1,2$, and $V^n(t) = V_1^n(t) + V_2^n(t)$, where $(L_l^n(t),\ t \geq 0)$ is the unfinished work process at queue l. In analogy with the queue-length process, $(V^n(t),\ t \geq 0)$ converges in distribution to an RBM $(V(t),\ t \geq 0)$ as $n \to \infty$, in this case with drift and diffusion coefficient

$$c_v = \lim_{n\to\infty} \sqrt{n}\,(\rho^n - 1), \qquad \rho^n = \rho_1^n + \rho_2^n, \tag{5.1}$$

$$\sigma_v^2 = \lambda_1\left(\sigma_{s1}^2 + \rho_1^2\sigma_{al}^2\right) + \lambda_2\left(\sigma_{s2}^2 + \rho_2^2\sigma_{a2}^2\right). \tag{5.2}$$

With the obvious extensions to conditions (2.1)–(2.4), Theorem 2.1 carries over to the unfinished work process in the general system:

$$\int_0^T f(V_l^n(t))\,dt \to_d \int_0^T \left(\int_0^1 f(uV(t))\,du \right) dt, \qquad l = 1,2,\ n \to \infty. \tag{5.3}$$

In terms of $\rho_l = \lambda_l/\mu_l$, $l = 1,2$, and the unfinished work, Theorem 4.1 on virtual waiting times becomes

$$\int_0^T f(Z_l^n(t))\,dt \to_d \int_0^T \int_0^1 f(uV(t)/\rho_l)\,du\,dt, \qquad l = 1,2, \tag{5.4}$$

as $n \to \infty$. Note that (4.33) and (4.34) still apply, in this case with $V(t)$ the unfinished work in the general, asymmetric system and with $\rho_l = \lambda_l/\mu_l$, $l = 1,2$.

The extension of our results to general $M > 2$ requires more effort. However, it is not difficult to see what the limit process for unfinished work should be when $M > 2$. In the general asymmetric system, consider the unfinished work process $(V_1(t), \dots, V_M(t))$ as the position of a particle moving in R_+^M. For a fixed, total unfinished work $v > 0$, the limit process has the particle cycling deterministically around a closed path in the hyperplane $V_1(t) + \cdots + V_M(t) = v$. Generalizing Figure 1, the path is piecewise linear with vertices at the coordinate hyperplanes; the vertices correspond to those times when the server has just finished emptying one queue and is starting on the next. Moving from one vertex to the next corresponds to serving a queue; during this time the queue being served empties at a fixed rate, while the other queues grow at fixed rates.

To determine the closed path, we compute its vertices as follows. Assume that the server visits queues $1,\dots,M$ in that order. Let the lth vertex correspond to the lth queue, $1 \le l \le M$, and suppose a cycle of the particle begins at vertex 1, when service to queue 1 is about to start and queue M is empty. Let α_l, $1 \le l \le M$, $\alpha_M = 0$, be the fraction of the total unfinished work v in queue l at the beginning of a cycle, and let τ_l denote the time spent by the server at queue l during a cycle. Since $\rho_1 + \cdots + \rho_M = 1$, the server spends a fraction ρ_l of its time at queue l, so $\rho_l = \tau_l/\tau$, where $\tau = \tau_1 + \cdots + \tau_M$. At queue l, work arrives at rate ρ_l and is completed at rate 1. Then, since $\alpha_l v$ is the amount of work that arrived at queue l since the server last departed from there, we have

$$\alpha_l v = \rho_l \sum_{l+1 \le k \le M} \tau_k = \tau\rho_l \sum_{l+1 \le k \le M} \rho_k. \tag{5.5}$$

However, $\sum_{1 \le l \le M-1} \alpha_l = 1$, so

$$\sum_{1 \le l \le M-1} \alpha_l = \frac{\tau}{v} \sum_{1 \le l \le M-1} \rho_l \sum_{l+1 \le k \le M} \rho_k = \frac{\tau}{v} \sum_{1 \le l < k \le M} \rho_l \rho_k = 1,$$

and hence

$$\tau = v \Big/ \sum_{1 \le j < k \le M} \rho_j \rho_k. \tag{5.6}$$

Substitution into (5.5) gives

$$\alpha_l = \frac{\rho_l \sum_{l+1 \le k \le M} \rho_k}{\sum_{1 \le j < k \le M} \rho_j \rho_k}. \tag{5.7}$$

Next, define α_{kl} as the fraction of the total unfinished work in queue l at the time the server starts serving queue k, $1 \le k,\ l \le M$. The kth vertex is $(\alpha_{k1},\dots,\alpha_{kM})v$. By definition of the α_l's, we have

$$\alpha_{kl} = \alpha_l + \rho_l \sum_{1 \le j \le k-1} \frac{\tau_j}{v} = \alpha_l + \frac{\tau}{v}\rho_l \sum_{1 \le j \le k-1} \rho_j, \qquad 1 \le k \le l, 1 \le l \le M, \tag{5.8}$$

and

$$\alpha_{kl} = \frac{\tau}{v}\rho_l \sum_{l+1 \le j \le k-1} \rho_j, \qquad l+1 \le k \le M, 1 \le l \le M. \tag{5.9}$$

By (5.6) and (5.7), we can rewrite (5.8) as

$$\alpha_{kl} = \frac{\tau}{v}\rho_l \left[\sum_{1 \le j \le k-1} \rho_j + \sum_{l+1 \le j \le M} \rho_j \right], \qquad 1 \le k \le l, 1 \le l \le M. \tag{5.10}$$

Together with (5.6), equations (5.9) and (5.10) determine the vertices of the path. To illustrate, we obtain the following vertices for $M = 3$:

$$(\alpha_{11}, \alpha_{12}, \alpha_{13})v = \left(\frac{\rho_1(\rho_2 + \rho_3)}{\rho_1 \rho_2 + \rho_1 \rho_3 + \rho_2 \rho_3}, \frac{\rho_2 \rho_3}{\rho_1 \rho_2 + \rho_1 \rho_3 + \rho_2 \rho_3}, 0 \right) v,$$

$$(\alpha_{21}, \alpha_{22}, \alpha_{23})v = \left(0, \frac{\rho_2(\rho_1 + \rho_3)}{\rho_1 \rho_2 + \rho_1 \rho_3 + \rho_2 \rho_3}, \frac{\rho_1 \rho_3}{\rho_1 \rho_2 + \rho_1 \rho_3 + \rho_2 \rho_3} \right) v,$$

$$(\alpha_{31}, \alpha_{32}, \alpha_{33})v = \left(\frac{\rho_1 \rho_2}{\rho_1 \rho_2 + \rho_1 \rho_3 + \rho_2 \rho_3}, 0, \frac{\rho_3(\rho_1 + \rho_2)}{\rho_1 \rho_2 + \rho_1 \rho_3 + \rho_2 \rho_3} \right) v.$$

As before, in the limit $n \to \infty$, the total unfinished work converges to RBM, but the particle speed tends to infinity. The waiting-time averaging principle of Section 1 [cf. (1.1)] is easily generalized. A random arrival again finds the particle positioned uniformly at random on the closed path. For convenience, let a cycle on this path begin when the server starts work at queue l. Let U, a random variable uniform on $[0, 1]$, denote the fraction of the current cycle that has transpired at the instant of a random arrival at queue l. Note that $0 \le U < \rho_l$ means that the server is at queue l. With the total unfinished work fixed at $V = v$, the work at queue l when the server starts his cycle there is $\rho_l(1 - \rho_l)\tau$ by (5.6) and (5.10). Then given $U = u$ and the unfinished work v, the conditional unfinished work at queue l seen by an arrival there is

$$\begin{aligned} V_l(u,v) &= \tau(1 - \rho_l)(\rho_l - u) \qquad (0 \le u \le \rho_l) \\ &= \tau\rho_l(u - \rho_l) \qquad (\rho_l \le u \le 1). \end{aligned} \tag{5.11}$$

The arrival must always wait $V_l(u, v)$, but if the server is not at queue l, the arrival must first wait for the current cycle to end. Then the conditional waiting time of an arrival at queue l is

$$\begin{aligned} Z_l &= V_l(u,v) \qquad (0 \le u \le \rho_l) \\ &= \tau(1 - u) + V_l(u,v) \qquad (\rho_l < u \le 1). \end{aligned} \tag{5.12}$$

Substitution for $V_l(u, v)$ and τ from (5.6) and (5.11) leads to the averaging principle

$$\begin{aligned} &\int_0^T f(Z_l^n(t))\,dt \\ &\to_d \int_0^T \left[\int_0^{\rho_l} f\left(V(t) \frac{1 - \rho_l}{\sum_{1 \le j < k \le M} \rho_j \rho_k} (\rho_l - u) \right) du \right. \\ &\qquad \left. + \int_{\rho_l}^1 f\left(V(t) \frac{1 - \rho_l}{\sum_{1 \le j < k \le M} \rho_j \rho_k} (1 + \rho_l - u) \right) du \right] dt, \quad l = 1,2. \end{aligned} \tag{5.13}$$

Changing variables of integration, (5.13) simplifies to

$$\int_0^T f(Z_l^n(t))\,dt \to_d \int_0^T \int_0^1 f\left(V(t) \frac{1 - \rho_l}{\sum_{1 \le j < k \le M} \rho_j \rho_k} u \right) du\,dt. \tag{5.14}$$

Note that (5.14) reduces to (5.4) when $M = 2$.

Sections 2–4 for the case $M = 2$ suggest strongly that the averaging principle (5.14) indeed holds. We leave the proof of this as an open problem, with the remark that a complete analysis for $M \geq 3$ must address a distinctly new issue, namely, convergence to a unique, deterministic particle motion for any fixed total load v. For $M \geq 3$ the particle can be off the closed path but still on the hyperplane $V_1(t) + \cdots + V_M(t) = v$. [This cannot happen when $M = 2$ because the hyperplane is itself the closed path $V_1(t) + V_2(t) = v$.] With V fixed, it is not hard to prove that in the limit process particles in such positions are attracted in an appropriate sense to a unique closed path, in particular the closed path determined by (5.9) and (5.10). The main difficulty is in constructing upper and lower bounds that allow us to handle fluctuations in the unfinished work.

Finally, an extension to nonzero switchover times is of obvious importance. A major problem is obtaining the diffusion limit for the total unfinished work, a trivial problem in the model of this paper. After obtaining the limit, which is a Bessel process, essentially the same program applies to give an averaging principle. An extension of Sections 2–4 to this case is planned for a forthcoming paper.

6. Waiting times. To illustrate the averaging principle in (5.14), this section gives heavy-traffic limits of normalized waiting (or sojourn) times. As in the usual model of stable queues in heavy traffic, we assume a negative drift $c_v < 0$. In this case, we have convergence to a stationary regime, namely, $V(t) \to \tilde{V}$, $t \to \infty$, where $\tilde{V}$ has the exponential distribution

$$P(\tilde{V} > z) = e^{-\gamma z}, \tag{6.1}$$

with $E[\tilde{V}] = 1/\gamma$, where $\gamma = 2|c_v|/\sigma_v^2$ and c_v and σ_v^2 are given by (5.1) and (5.2). We are interested in waiting times in statistical equilibrium, so for convenience in what follows, we take $(V(t),\ t \geq 0)$ as the stationary process. That is, $V(0) =_d \tilde{V}$. In this context, the averaging principle provides expressions for quantities averaged over all customers arriving during a time interval, as opposed to the same quantity for a single arrival.

We first calculate the moments of the stationary waiting time. We use (5.14) with $f(x) = x^k$. We obtain, with $r = \sum_{1 \leq j < k \leq M} \rho_j \rho_k$,

$$\begin{aligned} E\left[\frac{1}{T}\int_0^T (Z_l^n(t))^k\,dt\right] &\to E\left[\frac{1}{T}\int_0^T\int_0^1 \left[V(t)\frac{1-\rho_l}{r}\right]^k u^k\,du\,dt\right] \\ &= \left[\frac{1-\rho_l}{r}\right]^k \frac{1}{k+1} E[\tilde{V}^k] \\ &= \left[\frac{1-\rho_l}{r}\right]^k \frac{k!}{(k+1)\gamma^k}, \qquad 1 \leq l \leq M. \end{aligned} \tag{6.2}$$

Thus, the stationary average waiting time is $(1 - \rho_l)/2\gamma r$ and the variance is $5(1 - \rho_l)^2/(12\gamma^2 r^2)$.

We next calculate $G_l(z)$, which is the fraction of customers arriving (in steady state) to queue l that wait no more than z. This is a surrogate for the sojourn time distribution. To calculate $G_l(z)$ we take $f(x) = 1(x > z)$ in (5.14) (Theorem 2.1 can easily be extended to cover this f). We then obtain

$$\begin{aligned} &E\left[\frac{1}{T}\int_0^T 1(Z_l^n(t) > z)\,dt\right] \\ (6.3) \qquad &\to E\left[\frac{1}{T}\int_0^T\int_0^1 1\left(u\frac{(1-\rho_l)v(t)}{r} > z\right)du\,dt\right] \\ &\equiv 1 - G_l(z), \qquad 1 \le l \le M. \end{aligned}$$

By the assumed stationarity of $V(t)$, we can rewrite the right-hand side of (6.3) as

$$\int_0^\infty\left[\int_0^1 1\left(u > \frac{rz}{x(1-\rho_l)}\right)du\right]\gamma e^{-\gamma x}\,dx = \int_0^\infty\left[1 - \frac{rz}{x(1-\rho_l)}\right]^+\gamma e^{-\gamma x}\,dx.$$

A calculation then gives

$$(6.4)\quad 1 - G_l(z) = \exp\left(-\frac{\gamma rz}{1-\rho_l}\right) - \gamma\frac{r}{1-\rho_l}z\int_{rz/(1-\rho_l)}^\infty \frac{\exp(-\gamma x)}{x}\,dx.$$

In terms of the exponential integral $\mathrm{Ei}(x) = \int_{-\infty}^x (e^{-t}/t)\,dt$, we have

$$G_l(z) = 1 - \exp\left(-\frac{\gamma rz}{1-\rho_l}\right) - \gamma\frac{r}{(1-\rho_l)}z\,\mathrm{Ei}\left(-\gamma\frac{rz}{(1-\rho_l)}\right), \qquad 1 \le l \le M.$$

A simpler expression, accurate for large z, is available for the tail. Change variables in (6.4) to obtain

$$\begin{aligned} &1 - G_l(z) \\ &\quad = \exp\left(-\frac{\gamma rz}{1-\rho_l}\right)\left[1 - \int_0^\infty \frac{\exp(-y)}{1 - y/(\gamma(r/(1-\rho_l))z)}\,dy\right], \qquad 1 \le l \le M. \end{aligned}$$

Expanding $[1 - y/(\gamma(r/(1-\rho_l))z)]^{-1}$ and integrating term by term, we find

$$1 - G_l(z) = \frac{(1-\rho_l)\exp(-\gamma rz/(1-\rho_l))}{\gamma rz}\left[1 + O\left(\frac{1}{z}\right)\right], \qquad l = 1, 2.$$

It is interesting to compare the polling server with the first come–first served (FCFS) server. After a service completion in the latter system, the server always goes to the queue having the customer that has waited the longest, and serves that customer next. Waiting times in the FCFS system, which are distributed exponentially as in (6.1), can be expected to have a smaller variance. This is borne out by the above calculations. For example, with $M = 2$ and symmetric loading $\rho_1 = \rho_2 = 1/2$, waiting times under polling have a variance $5/3\gamma^2$, whereas under FCFS they have a variance $1/\gamma^2$.

To adapt our results to exact results for the special case of Poisson arrivals, we examine $\lim_{\rho\to 1}(1-\rho)E[\tilde{Z}_l(\rho)]$ and compare with a formula of

Sykes (1970) for $M = 2$. With the choice $n = 1/(1-\rho)^2$, we obtain $c_v = -1$ by (5.1). (The choice of n does not actually matter: the n's cancel out in the final expression.) In the special case of Poisson arrivals to each queue, we obtain $\sigma_{al}^2 = 1/\lambda_l^2$, and for the diffusion coefficient, $\sigma_v{}^2 = \lambda_1 b_1^{(2)} + \lambda_2 b_2^{(2)}$, where $b_l^{(i)}$ denotes the ith moment of the service times at queue l, $l = 1, 2$. Then by (6.2),

$$E\big[\tilde{Z}_l\big] = \frac{\lambda_1 b_1^{(2)} + \lambda_2 b_2^{(2)}}{4\rho_l}, \qquad l = 1,2,$$

which matches the result given by Sykes (1970).

Acknowledgment. Thanks to a helpful referee, the current version of Theorem 2.1 is more general than the authors' original version.

REFERENCES

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

BOXMA, O. J. and TAKAGI, H., EDS. (1992). Special issue on polling systems. *Queueing Syst.* **11** (1 and 2).

IGLEHART, D. L. and WHITT, W. (1970). Multiple channel queues in heavy traffic, I and II. *Adv. in Appl. Probab.* **2** 355–364.

JACOD, J. and SHIRYAEV, A. N. (1987). *Limit Theorems for Stochastic Processes.* Springer, New York.

KRICHAGINA, E. V., LIPTSER, R. SH. and PUHALSKII, A. A. (1988). Diffusion approximation for a system with a queue-dependent input and arbitrary service. *Theory Probab. Appl.* **33** 124–135.

PETROV, V. V. (1975). *Sums of Independent Random Variables.* Springer, New York.

SYKES, J. S. (1970). Simplified analysis of an alternating-priority queueing model with setup times. *Oper. Res.* **18** 1182–1192.

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

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

E. G. COFFMAN, JR.
M. I. REIMAN
AT & T BELL LABORATORIES
MURRAY HILL, NEW JERSEY 07974

A. A. PUHALSKII
INSTITUTE FOR PROBLEMS IN
INFORMATION TRANSMISSION
MOSCOW
RUSSIA 101447