The Annals of Applied Probability
1996, Vol. 6, No. 1, 1–47

A MULTICLASS CLOSED QUEUEING NETWORK WITH UNCONVENTIONAL HEAVY TRAFFIC BEHAVIOR

BY J. M. HARRISON AND R. J. WILLIAMS[1]

Stanford University and University of California, San Diego

We consider a multiclass closed queueing network model analogous to the open network models of Rybko and Stolyar and of Lu and Kumar. The closed network has two single-server stations and a fixed customer population of size n. Customers are routed in cyclic fashion through four distinct classes, two of which are served at each station, and each server uses a preemptive-resume priority discipline. The service time distribution for each customer class is exponential, and attention is focused on the critical case where all four classes have the same mean service time. Letting n approach infinity, we prove a heavy traffic limit theorem that is unconventional in three regards. First, in our heavy traffic scaling of both queue-length processes and cumulative idleness processes, time is compressed by a factor of n rather than the factor of n^2 occurring in conventional theory. Second, the spatial scaling applied to some components of the queue-length and idleness processes is that associated with the central limit theorem, but the scaling applied to other components is that associated with the law of large numbers. Thus, in the language of queueing theory, our heavy traffic limit theorem involves a mixture of Brownian scaling and fluid scaling. Finally, the limit process that we obtain is not an ordinary reflected Brownian motion, as in conventional heavy traffic theorems, although it is related to or derived from Brownian motion.

Contents

Received June 1995; revised November 1995.

[1] Research supported in part by NSF Grant GER-90-23335.

AMS 1991 *subject classifications.* 60K25, 60J70, 90B15, 90B22.

Key words and phrases. Closed multiclass queueing networks, heavy traffic theory, reflecting barrier, Brownian motion, $\mathbf{M}_1$ convergence.

1. Introduction.

This paper is part of a long-term research project on Brownian models of complex queueing networks. Such Brownian system models, or Brownian approximations, arise as heavy traffic limits of conventional queueing models after an appropriate scaling of time and state space. In all of the heavy traffic limit theorems that have been proved to date, the scaling that gives convergence to a Brownian limit is that associated with the central limit theorem (CLT), and the Brownian model that emerges as the heavy traffic limit is some kind of reflected Brownian motion (RBM). For open queueing networks, the current state of knowledge regarding heavy traffic limit theory is surveyed by Harrison and Nguyen [10], and Williams [22] provides an up-to-date review of mathematical theory for the associated Brownian system models. For a restrictive class of closed queueing networks, analogous results on Brownian approximations and heavy traffic limit theory were proved by Chen and Mandelbaum [4] and by Harrison, Williams and Chen [12], but overall, less is known about Brownian limits or Brownian approximations for closed queueing networks than for open ones.

The theory referred to in the previous paragraph is useful because the Brownian system model that one obtains as a heavy traffic approximation, although subtle and complex in its own right, is simpler in all important regards than the conventional network model it replaces. A key point is that reflected Brownian motions form a cohesive class of stochastic processes for which both general mathematical theory and general methods of numerical analysis are available.

How broadly applicable is the conventional heavy traffic framework, where CLT scaling of a multidimensional queue-length process gives weak convergence to an RBM under heavy traffic conditions? To be more precise, what are the limits of its applicability and, for queueing networks outside those limits, are there other kinds of heavy traffic theorems from which one can derive useful approximate system models? To shed some light on these important questions, we consider in this paper a simple network model for which the conventional heavy traffic framework is inadequate. For this model we prove a heavy traffic limit theorem that is unconventional in three respects. First, our theorem involves a milder scaling of time than what one sees in conventional theory. Even with this relatively mild compression of the time scale, a legitimately stochastic limit is obtained, which shows that the model under study here has a higher degree of intrinsic stochastic variability than network models previously studied. The second unconventional feature of our heavy traffic limit theorem is that different components of the stochastic processes under study are subjected to different spatial scalings. As a result, our limit theorem involves a mixture of CLT scaling with what queueing theorists call fluid scaling. Finally, the multidimensional stochastic process obtained as a limit in our heavy traffic theorem is not an ordinary RBM, although it is related to or derived from Brownian motion. Some components of our limit process exhibit the unbounded variation characteristic of Brownian paths, while other components have bounded variation (this is to be expected from the mixture of CLT scaling and fluid scaling). Sample paths of the limit process also exhibit jumps in certain components. The convergence for these components is relative to the Skorokhod $\mathbf{M}_1$ topology rather than the usual $\mathbf{J}_1$ topology on path space (see Section 2 for more details).

The model on which we focus is a multiclass closed queueing network first studied by Harrison and Nguyen [11]. It is precisely analogous to the open network models introduced by Rybko and Stolyar [18] and by Lu and Kumar [15], which have played an important role in the recent explosion of research on open network stability. It has been shown that these open networks may be unstable, depending on parameter values, even when each station has a traffic intensity parameter strictly less than 1. The subtle behavior observed in our closed network analog derives from the same underlying structure that creates the potential for such instability.

The paper is organized as follows. First, some notation and mathematical preliminaries are laid out in Section 2. The closed network model to be studied is introduced in Section 3. There we also review a key observation by Harrison and Nguyen [11] and identify a parameter combination that produces the most delicate system behavior. A heavy traffic limit theorem for that "critical case" is stated in Section 5, after a review of conventional heavy traffic theory in Section 4. Our unconventional heavy traffic limit theorem for the critical case is proved in Sections 6 through 8, with heavy reliance on a system representation that fully exploits the special structure of our model. To make the flow of logic in Sections 7 and 8 more transparent, the proofs of certain properties are isolated in Appendixes A and B. Throughout the paper, results that are labelled as propositions, lemmas, theorems or corollaries are numbered according to a single sequential scheme, for example, Corollary 5.2 is the result immediately following Theorem 5.1.

2. Notation and preliminaries. For each positive integer m, let D^m be the space of "Skorokhod paths" in $\mathbb{R}^m$ having time domain $\mathbb{R}_+ = [0, \infty)$. That is, D^m consists of all functions x: $[0, \infty) \to \mathbb{R}^m$ that are right continuous on $\mathbb{R}_+$ and have finite left limits on $(0, \infty)$. The subspace of D^m consisting only of continuous functions is denoted by C^m. When $m = 1$, we shall simply write D, C instead of D^m, C^m, respectively. At different points in this paper we consider D^m under both the Skorokhod $\mathbf{J}_1$ topology and the weaker $\mathbf{M}_1$ topology. The original reference for these topologies on the space of Skorokhod paths defined over $[0, 1]$ is [20]. For the extension to paths defined over $[0, \infty)$, see [21] and also [8] for the $\mathbf{J}_1$ topology. When either the $\mathbf{J}_1$ or $\mathbf{M}_1$ topology is relativized to C^m, it is the topology of uniform convergence on compact time intervals. We shall write u.o.c. as an abbreviation for *uniformly on compacts*, to indicate that a sequence of functions in D^m (or C^m) is converging uniformly on compact time intervals to a limit in D^m (or C^m).

We refer the reader to [20, 21, 8] for the precise definitions of the $\mathbf{J}_1$ and $\mathbf{M}_1$ topologies. Heuristically, convergence in these topologies may be described as follows. Consider a sequence $\{x_n\}$ in D^m that converges in the $\mathbf{J}_1$ or $\mathbf{M}_1$ topology to $x \in D^m$. Then for either topology, at a continuity point t of x, $x_n(t) \to x(t)$ as $n \to \infty$. The distinction between the topologies comes in convergence near jumps of x. In the case of $\mathbf{J}_1$ convergence, around the time of a jump of x, x_n must have a *single* jump that is close in location and magnitude to that of x. In the case of $\mathbf{M}_1$ convergence, around the time of a jump of x, x_n may have several jumps and the graph of x_n must be almost a "monotone staircase" which converges to the graph of x as $n \to \infty$. For certain components of the multidimensional queue-length, we shall be proving (weak) convergence of processes with many small jumps to a limit process in which these small jumps may coalesce to big jumps. For this the $\mathbf{M}_1$ topology will prove to be the appropriate topology. The space D^m with the $\mathbf{J}_1$ or the $\mathbf{M}_1$ topology is a Polish space (see [16, 21]) and so we shall be able to use the Skorokhod representation theorem (see [8], Theorem 3.1.8) to reduce many of our weak convergence arguments to ones involving almost sure convergence. For this, the following properties of path convergence will be useful.

PROPOSITION 2.1. (i) *Suppose $x_n \to x$ in either the $\mathbf{J}_1$ or $\mathbf{M}_1$ topology on D^m. If $x \in C^m$, then $x_n \to x$ u.o.c.*

(ii) *Suppose that $x_n \to x$ and $y_n \to y$ in the $\mathbf{J}_1$ (respectively, $\mathbf{M}_1$) topology on D^m. Then $x_n + y_n \to x + y$ in the $\mathbf{J}_1$ (respectively, $\mathbf{M}_1$ topology) if x and y have no points of discontinuity in common.*

(iii) *Suppose x_n and x are nonnegative, nondecreasing functions in D. Then $x_n \to x$ in the $\mathbf{M}_1$ topology if and only if* (a) $x_n(0) \to x(0)$ *as $n \to \infty$, and* (b) x_n *converges pointwise to x at a dense set of times.*

REMARK. In (iii), (a) and (b) may be replaced by "x_n converges to x at all continuity points of x" (this includes convergence at $t = 0$ by the right continuity of x).

PROOF OF PROPOSITION 2.1. For (i), see [20]; for (ii), see [16], Section III, Theorem 3.1; for (iii) and the Remark, see [21], Remark following Theorem 7.1, and [20], Section 2.4.1. □

Now the space D^m is equal as a set to the Cartesian product of m copies of D. However, the product topology on D^m, where each copy of D is endowed with the $\mathbf{J}_1$ (respectively $\mathbf{M}_1$) topology, is weaker than the $\mathbf{J}_1$ (respectively $\mathbf{M}_1$) topology on D^m (cf. [1] and [21], Section 4). In the sequel we shall need both the $\mathbf{J}_1$ or $\mathbf{M}_1$ topology on D^m and the product topology on D^m, where the copies of D in the product have either the $\mathbf{J}_1$ or $\mathbf{M}_1$ topology; we shall even allow some copies of D in the product to have the $\mathbf{J}_1$ topology and the remainder to have the $\mathbf{M}_1$ topology. Whenever we need to use a product topology on D^m, this will be clearly indicated. Otherwise, $\mathbf{J}_1$ or $\mathbf{M}_1$ convergence refers to the usual $\mathbf{J}_1$ or $\mathbf{M}_1$ topology on D^m.

For stochastic processes $X_1, X_2, \ldots, X$ whose paths lie almost surely in D^m, we write "$X_n \Rightarrow X$ in the $\mathbf{J}_1$ topology" to mean that the probability measures induced by the X_n on D^m endowed with the $\mathbf{J}_1$ topology converge weakly to the probability measure induced on D^m by X; this same state of affairs may be expressed by the statement "X_n converges weakly in the $\mathbf{J}_1$ topology to X as $n \to \infty$." Weak convergence under the $\mathbf{M}_1$ topology is expressed similarly. On the other hand, when D^m is to be considered as a product of m copies of D, each with the $\mathbf{J}_1$ topology, we shall write "$X_n \Rightarrow X$ with the product topology on D^m, where each copy of D has the $\mathbf{J}_1$ topology." Similar terminology will be used when we have a product of m copies of D, each with the $\mathbf{M}_1$ topology or with some mixture of the $\mathbf{M}_1$ and $\mathbf{J}_1$ topologies (see Theorem 5.1).

Let D_0 be the subspace of D consisting of those functions $x \in D$ with initial value $x(0) \in [0,1]$. Let D_0^f denote the subspace of D_0 consisting of those functions in D_0 which jump at most finitely many times in any compact time interval.

The following proposition serves to define and characterize the two-sided reflection mapping (η_1, η_2, ρ): $D_0^f \to D^3$, which is also called by Harrison [9] the two-sided regulator. Critical continuity and measurability properties of this mapping are stated in Propositions 2.3 and 2.4. These three propositions can be obtained from the results in Chen and Mandelbaum ([3]; see Proposition 2.4, Theorem 2.5 and the Remark following it and Theorem 2.6) by first performing a linear transformation of the unit interval $[0,1]$ to the line segment $\{x = (x_1, x_2) \in \mathbb{R}^2\colon x_1 + x_2 = 1\}$.

PROPOSITION 2.2. *For each $x \in D_0^f$ there is a unique triple $(y_1, y_2, z) \in D^3$ satisfying*

$$z(t) = x(t) + y_1(t) - y_2(t), \qquad t \ge 0, \tag{2.1}$$

$$0 \le z(t) \le 1, \qquad t \ge 0, \tag{2.2}$$

$$y_1 \text{ and } y_2 \text{ are nondecreasing with } y_1(0) = y_2(0) = 0, \tag{2.3}$$

$$z(t) = 0 \text{ at every time } t \ge 0 \text{ that is a point of increase for } y_1 \tag{2.4}$$

and

$$z(t) = 1 \text{ at every time } t \ge 0 \text{ that is a point of increase for } y_2. \tag{2.5}$$

Moreover, y_1 and y_2 are the least *functions satisfying* (2.1)–(2.3), *in the following sense: If (y_1', y_2', z') is another triple satisfying* (2.1)–(2.3), *then $y_1(t) \le y_1'(t)$ and $y_2(t) \le y_2'(t)$ for all $t \ge 0$.*

DEFINITION. Given $x \in D_0^f$, let $\eta_1(x) = y_1$, $\eta_2(x) = y_2$ and $\rho(x) = z$, where (y_1, y_2, z) is the unique solution of (2.1)–(2.5).

PROPOSITION 2.3. *If $\{x_n\}$ is a sequence in D_0^f which converges u.o.c. to $x \in C$, then $\{(\eta_1, \eta_2, \rho)(x_n)\}$ converges u.o.c. to $(\eta_1, \eta_2, \rho)(x) \in C^3$ as $n \to \infty$.*

PROPOSITION 2.4. *If X is a one-dimensional stochastic process that has continuous paths* (*respectively, paths locally of bounded variation in D_0^f*) *starting in $[0,1]$, then $(\eta_1, \eta_2, \rho)(X)$ is a continuous* (*respectively, locally of bounded variation*) *stochastic process that is adapted to X.*

As usual, given a Borel set $A \subset \mathbb{R}$, we define the indicator function 1_A: $\mathbb{R} \to \{0, 1\}$ by setting $1_A(x) = 1$ if $x \in A$ and $= 0$ otherwise.

3. The closed queueing network. Consider a closed system with two single-server stations and n customers who circulate perpetually with the deterministic routing pictured in Figure 1. A customer cycle consists of four services at stations 1, 2, 2 and 1 again, in that order. Customers that are waiting for or undergoing the kth service of their cycle will be called class k customers $(k = 1, 2, 3, 4)$. All service times are independent and class k service times are assumed to have an exponential distribution with mean $m_k > 0$. Finally, each server follows a preemptive-resume priority discipline, as shown in Figure 1.

Let $Q_k(t)$ denote the number of class k customers existing at time t, calling this the queue-length for class k, and define a four-dimensional queue-length process $Q = \{Q(t),\ t \geq 0\}$ in the obvious way. With the assumptions enunciated above, Q is a continuous time Markov chain with finite state space. The following proposition is due to Harrison and Nguyen [11]. Its proof is included for completeness.

PROPOSITION 3.1. *Given any initial queue-length vector* $Q(0)$, *let*

$$\tau = \inf\{t \geq 0 \colon Q_2(t) = 0 \text{ or } Q_4(t) = 0\}.$$

Then, almost surely, $\tau < \infty$ *and* $Q_2(t)Q_4(t) = 0$ *for all* $t \geq \tau$.

REMARK. In words, this says that after a finite initial time interval, the two servers will never again have priority work to do at the same time.

PROOF OF PROPOSITION 3.1. Suppose that $Q_2(0) = i > 0$ and $Q_4(0) = j > 0$ (that is, each server has priority work to do initially). Let $S_2(i)$ be the sum of the first i class 2 service times and define $S_4(j)$ similarly. Now $\tau = \min\{S_2(i), S_4(j)\}$. During the interval $[0, \tau]$ no effort is devoted to service of the nonpriority customers in classes 1 and 3, so the priority queue-lengths Q_2 and Q_4 are nonincreasing over that interval. It follows that τ is the stopping time identified in the statement of the proposition, and clearly $E(\tau) < \infty$.

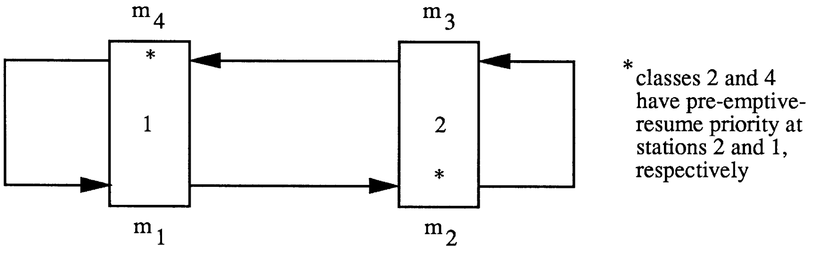
m4
m3
*
1
2
*
m1
m2
*classes 2 and 4 have pre-emptive-resume priority at stations 2 and 1, respectively

FIG. 1. *A multiclass closed queueing network with priority service.*

Let us now consider the evolution of the system beginning from a state in which one or both of the priority queues (that is, the queues for classes 2 and 4) are empty. For the sake of concreteness, assume $Q_2(0) > 0$ and $Q_4(0) = 0$. Then server 2 will work on class 2 (priority) customers up until the first time σ at which $Q_2(\sigma) = 0$, and during the interval $[0, \sigma]$ no effort will be devoted to service of the nonpriority customers in class 3, so no new customers of class 4 can be created. Thus we have $Q_2(\sigma) = Q_4(\sigma) = 0$. When the next service is completed at some time $t > \sigma$, the system will return to a condition in which one but not both of the priority queues (those for classes 2 and 4) is empty, and now the argument repeats: when just one of the priority classes is being served, no new customers of the other priority class can be created and so $Q_2(\cdot)Q_4(\cdot) = 0$. □

Proposition 3.1 shows that some states of the Markov chain Q are transient, namely, those with $Q_2 > 0$ and $Q_4 > 0$. To avoid trivial complications, we assume hereafter that $Q_2(0)Q_4(0) = 0$. The queue-length process Q is then effectively two dimensional, because

$$(3.1) \qquad Q_2(t)Q_4(t) = 0 \quad \text{and} \quad \sum_{k=1}^{4} Q_k(t) = n \quad \text{for all } t \geq 0.$$

For our purposes a particularly convenient two-dimensional representation is the following. Let

$$(3.2) \qquad V_1(t) = Q_2(t) - Q_4(t) \quad \text{and} \quad V_2(t) = Q_1(t) + Q_2(t).$$

From (3.1) we see that the four-vector $Q(t)$ can be recovered from the two-vector $V(t)$ by means of the identities

$$(3.3) \qquad Q_2(t) = [V_1(t)]^+,$$

$$(3.4) \qquad Q_4(t) = [V_1(t)]^-,$$

$$(3.5) \qquad Q_1(t) = V_2(t) - Q_2(t)$$

and

$$(3.6) \qquad Q_3(t) = n - [Q_1(t) + Q_2(t) + Q_4(t)].$$

Thus V is also a Markov chain. Its state space and transition structure are pictured in Figure 2, and it is easy to write out the intensity parameters for the various transitions pictured (each transition occurs at rate $\mu_k = 1/m_k$ for some k).

To close this section we shall summarize some implications of Proposition 3.1 with regard to the long-run system throughput rate. Let $T_k(t)$ denote the total amount of time devoted to service of class k (by whichever server handles that class) over the interval $[0, t]$. Assuming as before that $Q_2(0)Q_4(0) = 0$, it follows from Proposition 3.1 that

$$(3.7) \qquad T_2(t) + T_4(t) \leq t, \qquad t \geq 0.$$

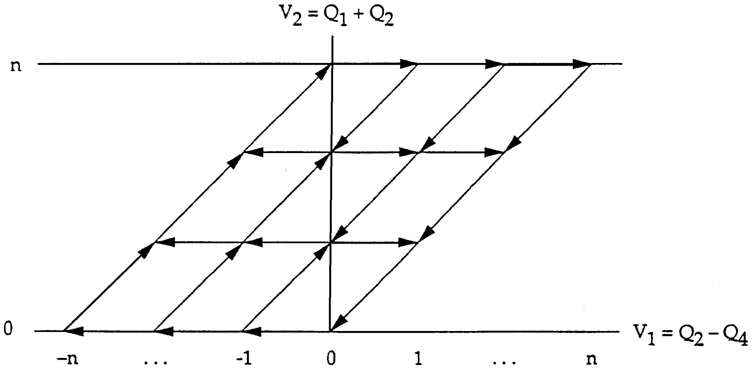

FIG. 2. *Transition structure of the Markov chain V (for $n = 3$).*

For each class k there is a constant θ_k (it is easy to express these constants in terms of the stationary distribution of the Markov chain Q) such that $E[T_k(t)] \sim \theta_k t$ as $t \to \infty$, independent of initial conditions. Moreover, from the simple cyclic routing pictured in Figure 1 it follows that (recall $\mu_k \equiv 1/m_k$)

$$\mu_1\theta_1 = \mu_2\theta_2 = \mu_3\theta_3 = \mu_4\theta_4 = \lambda \tag{3.8}$$

for some constant λ called the *system throughput rate*, and (3.8) can be written equivalently as

$$\theta_k = \lambda m_k \quad \text{for each class } k = 1,2,3,4. \tag{3.9}$$

Obviously, $\theta_1 + \theta_4 \leq 1$ because classes 1 and 4 are both served at station 1, and similarly $\theta_2 + \theta_3 \leq 1$. However, (3.7) further implies that $\theta_2 + \theta_4 \leq 1$, and combining these three inequalities with (3.9) gives

$$\lambda \leq \lambda^* \equiv \min\{(m_1 + m_4)^{-1}, (m_2 + m_3)^{-1}, (m_2 + m_4)^{-1}\}. \tag{3.10}$$

The upper bound λ^* in (3.10) is independent of n and the arguments in Harrison and Nguyen [11] can be extended to show that λ does in fact approach this bound as $n \to \infty$.

Obviously, $(m_1 + m_4)^{-1}$ represents the maximum rate at which server 1 can process incoming customers (or equivalently, the server's average processing rate if never starved for work) and $(m_2 + m_3)^{-1}$ is the analogous quantity for server 2. To get an interesting heavy traffic limit theorem, we shall assume hereafter that the two servers' maximum processing rates are equal, calling this a *balanced loading condition*. By choosing units appropriately, we can express the balanced loading condition as

$$m_1 + m_4 = m_2 + m_3 = 1. \tag{3.11}$$

Combining (3.8) with assumption (3.11), one has

$$\theta_1 + \theta_4 = \theta_2 + \theta_3 = \lambda, \tag{3.12}$$

which means that the long-run utilization rate for both server 1 and server 2 is equal to the system throughput rate λ.

An unsuspecting analyst might paraphrase (3.11) by saying that server 1 and server 2 are tied for bottleneck status, and expect that $\lambda \to 1$ as $n \to \infty$. However, (3.10) shows that a *hidden bottleneck* exists if $m_2 + m_4 > 1$, and in that circumstance the long-run utilization rate remains bounded away from 1 (or equivalently, the long-run idleness rate remains bounded away from 0) for both server 1 and server 2 as $n \to \infty$. Given our balanced loading condition (3.11), one has $m_2 + m_4 > 1$ if and only if

$$m_1 < m_2 \quad \text{and} \quad m_3 < m_4. \tag{3.13}$$

That is, the hidden bottleneck emerges in our balanced closed network when the nonpriority service operations are faster on average than the priority service operations which they precede. Readers who wish to understand exactly how the hidden bottleneck affects system dynamics are referred to the brilliant analysis by Dai and Weiss [6] of the open network models of Lu and Kumar and of Rybko and Stolyar.

Maintaining the balanced loading assumption (3.11) and motivated by the discussion above, we shall identify the following parameter ranges later in this paper:

subcritical case:

$$m_2 + m_4 < 1 \quad (\text{that is, } m_1 > m_2 \text{ and } m_3 > m_4).$$

critical case:

$$m_2 + m_4 = 1 \quad (\text{that is, } m_1 = m_2 \text{ and } m_3 = m_4).$$

supercritical case:

$$m_2 + m_4 > 1 \quad (\text{that is, } m_1 < m_2 \text{ and } m_3 < m_4).$$

As discussed above, neither server is able to approach full utilization as $n \to \infty$ in the supercritical case, where a hidden bottleneck emerges as the unique limiting factor on server utilization or system throughput. In the critical and subcritical cases, full utilization *is* approached (i.e., $\lambda \to 1$) as $n \to \infty$, but it will be shown later that the hidden bottleneck still asserts itself in a certain sense as $n \to \infty$ in the critical case.

4. Heavy traffic behavior when priorities are reversed. A closed queueing network is said to be "in heavy traffic" if its population size n is large, and a "heavy traffic limit" involves letting $n \to \infty$. The central purpose of this paper is to state and prove a heavy traffic limit theorem for the critical case ($m_1 = m_2$ and $m_3 = m_4$) identified at the end of Section 3. To set the stage, we describe in this section a "conventional" heavy traffic limit theorem that provides a useful point of comparison for our main result.

The conventional heavy traffic limit theorem involves a sequence of closed networks indexed by $n = 1, 2, \dots,$ each having the structure described in Section 3 *except that the service priorities at each station are reversed.* That is, for purposes of this section only, let us assume that class 1 has preemptive-resume priority at station 1 and class 3 has preemptive-resume priority at station 2. Thus each server gives preference to exiting customers over entering customers. The nth system has a population of size n and the mean service times m_k are fixed (not depending on n) and are assumed to satisfy the balanced loading condition (3.11). Denoting by $\{Q^n(t),\ t \geq 0\}$ the four-dimensional queue-length process associated with the nth system, let us assume the convenient initial conditions

$$Q^n(0) = (0, 0, 0, n) \quad \text{for all } n. \tag{4.1}$$

With the reversed priorities assumed in this section, there can never be more than one customer of class 1, because each service of a low-priority class 4 customer at station 1 is followed immediately by the high-priority class 1 service of that same customer. Similarly, each service of a low-priority class 2 customer at station 2 is followed immediately by the high-priority class 3 service of that same customer.

Given this state of affairs, our original multiclass closed network is equivalent to a closed network with a single class served at each station. In the equivalent single-class network customers visit stations 1 and 2 alternately, each service at station 1 is distributed as the sum of a class 4 and a class 1 service, and each service at station 2 is distributed as the sum of a class 2 and a class 3 service. According to (3.11), the expected total service time at each station is 1 and the total service time at each station obviously has finite variance (recall that each class's service time distribution was assumed to be exponential).

Chen and Mandelbaum [4] proved a heavy traffic limit theorem for single-class closed networks, which specializes to the case at hand as follows. First define a sequence of four-dimensional scaled queue-length processes $\tilde{Q}^n(t)$ via

$$\tilde{Q}^n(t) = \frac{1}{n} Q^n(n^2 t), \qquad t \geq 0. \tag{4.2}$$

The scaling of queue lengths by a factor of n is entirely natural, since $\tilde{Q}^n_k(\cdot)$ expresses the class k queue length as a fraction of the total population, and then CLT scaling requires a corresponding compression of the time scale by a factor of n^2. By analogy with (4.2) we define a sequence of two-dimensional scaled idleness processes

$$\tilde{I}^n(t) = \frac{1}{n} I^n(n^2 t), \qquad t \geq 0, \tag{4.3}$$

where $I^n(t) = (I^n_1(t), I^n_2(t))$ and $I^n_j(t)$ is the cumulative idleness suffered by server j up to time t in the nth system. From the Chen–Mandelbaum limit theorem one easily deduces that, as $n \to \infty$,

$$\left(\tilde{Q}^n, \tilde{I}^n\right) \Rightarrow (Q^*, I^*) \quad \text{in the } \mathbf{J}_1 \text{ topology}. \tag{4.4}$$

The six-dimensional limit process (Q^*, I^*) is given by

$$(Q^*, I^*) = (0, Z^*, 0, 1 - Z^*, Y^*_1, Y^*_2), \tag{4.5}$$

where Z^* is a one-dimensional reflected Brownian motion on the interval $[0,1]$ with zero drift and a certain variance parameter $\sigma^2 > 0$ (computable from the mean service times m_k), Y_1^* is a multiple of the local time process associated with the boundary $Z^* = 1$ and Y_2^* is a multiple of the local time process associated with the boundary $Z^* = 0$. Given our assumed initial condition (4.1), the limit process (Q^*, I^*) has initial state

$$(Q^*(0), I^*(0)) = (0,0,0,1,0,0). \tag{4.6}$$

In applications of closed network theory, greatest interest usually attaches to questions of system throughput or, equivalently, to questions of server idleness, and the heavy traffic limit theorem (4.4) has a great deal to say in this regard. First, assuming the necessary uniform integrability, from (4.4) it follows that

$$\begin{aligned} E\big[\tilde{I}_1^n(t)\big] &\equiv \frac{1}{n} E\big[I_1^n(n^2 t)\big] \\ &\to E\big[I_1^*(t)\big] \quad \text{as } n \to \infty \quad \text{for each fixed } t > 0. \end{aligned} \tag{4.7}$$

Now let us define

$$\gamma^n(t) = \frac{1}{t} E\big[I_1^n(t)\big] \quad \text{and} \quad \gamma^*(t) = \frac{1}{t} E\big[I_1^*(t)\big],$$

so that $\gamma^n(t)$ is the average idleness rate over $[0,t]$ for server 1 in the nth queueing network and $\gamma^*(t)$ is an analogous quantity for the limiting Brownian system model. Then (4.7) can be rewritten as

$$\lim_{n\to\infty} \big[n\gamma^n(n^2 t)\big] = \gamma^*(t) \quad \text{for fixed } t > 0. \tag{4.8}$$

A long-run average idleness rate $\gamma^n(\infty) \equiv \lim_{t\to\infty} \gamma^n(t)$ is known to exist for each system n in our sequence (this limit is independent of the particular initial conditions assumed here) and the limit $\gamma^*(\infty)$ is also known to exist (it too is independent of initial conditions). If an exchange of limits can be justified, then (4.8) will give

$$\gamma^n(\infty) \sim \frac{1}{n}\gamma^*(\infty) \quad \text{as } n \to \infty, \tag{4.9}$$

thus quantifying the rate at which long-run server idleness vanishes as the system's population size n grows large. Of course, (4.7) further suggests that a time span which is large compared to n^2 is necessary for the long-run average to be approached, so one must be careful about facile use of (4.9).

Again as a point of comparison for results developed later, let us define scaled queue-length processes $\bar{Q}^n$ and scaled idleness processes $\bar{I}^n$ via

$$\bar{Q}^n(t) = \frac{1}{n} Q^n(nt), \qquad t \ge 0, \tag{4.10}$$

and

$$\bar{I}^n(t) = \frac{1}{n} I^n(nt), \qquad t \ge 0. \tag{4.11}$$

The scaling embodied in (4.10) and (4.11), wherein the space and time scales are compressed by the same factor, is that associated with the law of large numbers, and in queueing theory it is often called “fluid scaling.” Given our balanced loading condition (3.11) and initial condition (4.1), one has from the Chen–Mandelbaum theory [4] that

$$(4.12) \qquad \left(\overline{Q}^n, \bar{I}^n\right) \Rightarrow (0,0,0,1,0,0) \quad \text{in the } \mathbf{J}_1 \text{ topology as } n \to \infty,$$

where the right-hand side of (4.12) is understood to mean a constant process whose fourth component has value 1 at all times $t \geq 0$ and so forth. Roughly speaking, (4.12) says that if n is large and we begin with all customers in class 4, then over a time span of order n, changes in the queue-length vector and cumulative idleness will both be $o(n)$.

Comparing the CLT or Brownian scaling in (4.2) and (4.3) with the fluid scaling in (4.10) and (4.11), we see that they rescale space variables in the same way, but Brownian scaling involves a more severe compression of the time scale, leading us to observe queue-lengths and cumulative idleness over time spans of order n^2 rather than order n. Thus it is plausible that processes which appear to be nearly constant for large n under fluid scaling would have significant stochastic variability when observed on the Brownian time scale.

To recapitulate, we have considered in this section the simple closed network model that one obtains when the priority rankings originally specified in Figure 1 are reversed. We have described a “conventional” heavy traffic limit theorem for that simple network, assuming that its mean service times m_k satisfy the balanced loading condition (3.11), but imposing no further restrictions on them. Consider now the original priority rankings specified in Figure 1, maintaining the balanced loading assumption ($m_1 + m_4 = m_2 + m_3 = 1$). For the subcritical case identified at the end of Section 3 ($m_1 > m_2$ and $m_3 > m_4$) we conjecture that the conventional limit theorem (4.4) holds after just trivial changes (e.g., it is the high-priority scaled queue-length processes, now for classes 2 and 4, that vanish in the limit). For the critical case ($m_1 = m_2$ and $m_3 = m_4$) it will now be shown that a very different sort of system behavior emerges in the heavy traffic limit. In the supercritical case ($m_1 < m_2$ and $m_3 < m_4$) one presumably obtains yet another mode of system behavior as $n \to \infty$, but we shall not even venture a guess at this time as to the form that branch of the heavy traffic theory will take.

5. The heavy traffic limit theorem (critical case). Let us return now to the closed priority network pictured in Figure 1, analysis of which was begun in Section 3. As explained there, the critical parameter combination is that where $m_1 = m_2$ and $m_3 = m_4$. For ease of exposition, we shall further specialize to the case where all mean service times are equal and then the unit of time can be chosen so that

$$(5.1) \qquad m_1 = m_2 = m_3 = m_4 = \tfrac{1}{2}.$$

With (5.1) assumed hereafter, we consider a sequence of networks with $n \to \infty$, using a superscript n to denote a process associated with the nth system. Maintaining the notation introduced in Sections 3 and 4, let us define scaled queue-length processes $\hat{Q}^n$ and scaled cumulative idleness processes $\hat{I}^n$ as

$$\hat{Q}_1^n(t) = \frac{1}{n} Q_1^n(nt) \quad \text{and} \quad \hat{Q}_3^n(t) = \frac{1}{n} Q_3^n(nt), \tag{5.2}$$

$$\hat{Q}_2^n(t) = \frac{1}{\sqrt{n}} Q_2^n(nt) \quad \text{and} \quad \hat{Q}_4^n(t) = \frac{1}{\sqrt{n}} Q_4^n(nt), \tag{5.3}$$

$$\hat{I}_1^n(t) = \frac{1}{\sqrt{n}} I_1^n(nt) \quad \text{and} \quad \hat{I}_2^n(t) = \frac{1}{\sqrt{n}} I_2^n(nt). \tag{5.4}$$

In the current context it will be convenient to assume that all customers in each system are initially waiting for class 1 service. Thus

$$\big(\hat{Q}^n(0), \hat{I}^n(0)\big) = (1,0,0,0,0,0) \quad \text{for all } n. \tag{5.5}$$

THEOREM 5.1. *Let the space D^6, where $(\hat{Q}^n, \hat{I}^n)$ takes its values, be endowed with the product topology for the product of six copies of D, where the first and third copies of D are endowed with the $\mathbf{J}_1$ topology and the other copies are endowed with the $\mathbf{M}_1$ topology. That is, think of D^6 with this topology as the product*:

$$(D, \mathbf{J}_1) \times (D, \mathbf{M}_1) \times (D, \mathbf{J}_1) \times (D, \mathbf{M}_1) \times (D, \mathbf{M}_1) \times (D, \mathbf{M}_1).$$

Then, as $n \to \infty$,

$$\big(\hat{Q}^n, \hat{I}^n\big) \Rightarrow (Q^*, I^*) \quad \textit{with the product topology on } D^6 \textit{ specified above},$$

where (Q^*, I^*) is the stochastic process defined immediately below.

REMARK. In fact, the convergence in the product topology can be refined slightly. For this, see the comments in the second to the last paragraph of Section 8.

Actually, the weak limit (Q^*, I^*) in Theorem 5.1 is defined in terms of a Markov process U^* whose precise mathematical construction will be delayed until Section 7. Informally, however, it is quite easy to explain how U^* behaves. First, its state space is the infinite strip pictured in Figure 3. Second, when $0 < U_2^* < 1$, the horizontal component U_1^* evolves as a Brownian motion with drift parameter equal to zero and variance parameter equal to 4. Third, U_2^* moves upward at the deterministic rate 2 on the left side of the strip and moves downward at rate 2 on the right side. Finally, when either the upper left or lower right portion of the strip's boundary is hit, there is an immediate jump to $U_1^* = 0$ (see Figure 3). Thus U_1^* can be decomposed as

$$U_1^*(t) = 2W^*(t) - J_1^*(t) + J_2^*(t), \qquad t \ge 0, \tag{5.6}$$

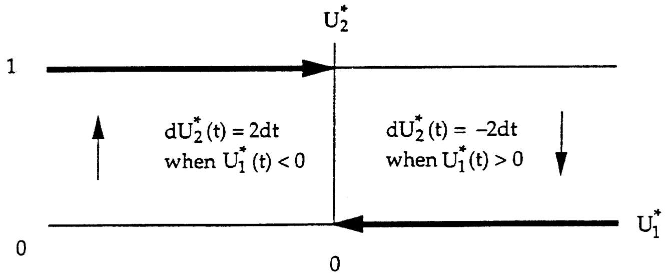
U_2^*
1
$dU_2^*(t) = 2dt$ when $U_1^*(t) < 0$
$dU_2^*(t) = -2dt$ when $U_1^*(t) > 0$
0
U_1^*
0

FIG. 3. *The Markov process U^*.*

where W^* is a standard Brownian motion, J_1^* is a nondecreasing process associated with the lower right boundary segment and J_2^* is a nondecreasing process associated with the upper left boundary segment. Furthermore, U_2^* has the decomposition (see Section 7)

$$U_2^*(t) = 1 + 2\int_0^t 1_{(-\infty,0)}(U_1^*(s))\,ds - 2\int_0^t 1_{(0,\infty)}(U_1^*(s))\,ds. \tag{5.7}$$

The limit (Q^*, I^*) is defined in terms of U^* as

$$Q_1^*(t) = U_2^*(t) \quad \text{and} \quad Q_3^*(t) = 1 - U_2^*(t), \tag{5.8}$$

$$Q_2^*(t) = [U_1^*(t)]^+ \quad \text{and} \quad Q_4^*(t) = [U_1^*(t)]^-, \tag{5.9}$$

$$I_1^*(t) = \tfrac{1}{2}J_1^*(t) \quad \text{and} \quad I_2^*(t) = \tfrac{1}{2}J_2^*(t). \tag{5.10}$$

After U^* has been defined precisely in Section 7, Theorem 5.1 will be proved in Section 8. The remainder of this section is devoted to a discussion of the theorem's intuitive content, particularly its differences from the conventional heavy traffic limit theory sketched earlier in Section 4.

All six of the processes defined by (5.8)–(5.10) are nondeterministic. If we define fluid-scaled processes $\overline{Q}^n$ and $\bar{I}^n$ via

$$\overline{Q}^n(t) = \frac{1}{n}Q^n(nt) \quad \text{and} \quad \bar{I}^n(t) = \frac{1}{n}I^n(nt), \qquad t \ge 0, \tag{5.11}$$

as in Section 4, then the following corollary is immediate from Theorem 5.1.

COROLLARY 5.2. *We have*

$$\left(\overline{Q}^n, \bar{I}^n\right) \Rightarrow (Q_1^*, 0, Q_3^*, 0, 0, 0) \quad \textit{in the } \mathbf{J}_1 \textit{ topology as } n \to \infty. \tag{5.12}$$

Thus, both priority queue-length processes and both cumulative idleness processes are asymptotically null under fluid scaling, as in conventional heavy traffic theory, but in the critical case fluid scaling gives a stochastic limit for the nonpriority queue lengths. By adopting the more delicate scaling (5.3) and (5.4) for priority queue-lengths and cumulative idleness, respectively, we have obtained a refinement of (5.12). Moreover, the limits Q_2^*, Q_4^*

under the milder scaling are needed to describe the stochastic behavior of the fluid limits Q_1^*, Q_3^*.

Given our emphasis on the unconventional heavy traffic scaling in Theorem 5.1, the following question is natural: for any one of the queue-length processes or cumulative idleness processes, is there another scaling of time and state space, different from that used in (5.2)–(5.4), that also gives weak convergence to a nondeterministic limit? A preliminary investigation has shown this question to be surprisingly subtle, and having identified it as an attractive topic for future research, we shall make no further comment on the matter in this paper.

An appealing feature of heavy traffic theory for closed queueing networks is that the large parameter n used for purposes of scaling is the total population size, a quantity with intrinsic significance. (In contrast, heavy traffic theorems for open networks are customarily stated in terms of a large parameter n that quantifies the rate of convergence in a sequence of system parameters hypothesized by the mathematical analyst, which makes physical interpretation of the limit theory difficult.) In the current context we have found that for a large population size n, stochastic variability in queue-lengths and cumulative idleness processes can be observed over time spans of order n, and over such time spans the variability in some processes is of order n, while for others it is of order $\sqrt{n}$. The conventional Brownian limit theorem (4.4) described earlier says that for large n one must observe the network for longer time spans of order n^2 to see significant stochastic variability, and that over such time spans the variability in both queue-length and cumulative idleness processes is of order n.

As stated earlier in Section 4, questions involving system throughput are usually of greatest interest in applications of closed queueing network models, and these can be equivalently recast as questions about cumulative server idleness. Again, assuming the necessary uniform integrability, as an analog of the conventional heavy traffic result (4.7) for expected cumulative idleness, we obtain from Theorem 5.1 that for all but countably many $t > 0$,

$$E\left[\hat{I}_1^n(t)\right] \equiv \frac{1}{\sqrt{n}} E\left[I_1^n(nt)\right] \to E\left[I_1^*(t)\right] \quad \text{as } n \to \infty. \tag{5.13}$$

[The exceptional set of t's where (5.13) may fail to hold consists of those at most countably many t at which I_1^* has a jump with positive probability (cf. [1], page 124).] Now let the average idleness rates $\gamma^n(t)$ and $\gamma^*(t)$ be defined as in Section 4, meaning that

$$\gamma^n(t) = \frac{1}{t} E\left[I_1^n(t)\right] \quad \text{and} \quad \gamma^*(t) = \frac{1}{t} E\left[I_1^*(t)\right].$$

Then (5.13) can be rewritten as

$$\lim_{n \to \infty} \left[\sqrt{n}\, \gamma^n(nt)\right] = \gamma^*(t). \tag{5.14}$$

Assuming that the limits $\gamma^n(\infty)$ and $\gamma^*(\infty)$ exist and that an exchange of limits can be justified, we arrive at the following analog of (4.9):

$$\gamma^n(\infty) \sim \frac{1}{\sqrt{n}}\gamma^*(\infty) \quad \text{as } n \to \infty. \tag{5.15}$$

For large values of n, of course, a long-run idleness rate of order $n^{-1/2}$ is much less favorable than the idleness rate of order n^{-1} predicted by (4.9) as a part of conventional heavy traffic theory. If, for example, we simply set $\gamma^*(\infty) = 1$ in both (4.9) and (5.15), the former estimates long-run server utilization at 99% with a population of size $n = 100$, while the latter estimates utilization of 90% for the same case.

6. A convenient representation of the queueing process. Consider a single closed queueing network of the type described in Section 3, with the population size n fixed throughout this section. Restricting attention to the critical case with $m_k = \frac{1}{2}$ for all four customer classes k, we can construct all stochastic processes of interest from two independent Poisson processes as follows.

Let $A_1 = \{A_1(t),\ t \ge 0\}$ and $A_2 = \{A_2(t),\ t \ge 0\}$ be independent, right continuous Poisson processes, each with arrival rate (or intensity parameter) 2 and with $A_1(0) = A_2(0) = 0$, defined on some probability space $(\Omega, \mathscr{F}, P)$. One may interpret A_1 and A_2 as the cumulative potential service processes at stations 1 and 2, respectively. Defining the two-dimensional process $A = (A_1, A_2)$, let $\{\mathscr{F}_t,\ t \ge 0\}$ be the filtration generated by A, satisfying the usual conditions. Defining the two-dimensional vector $e = (1, 1)$, it will be useful to recall the martingale characterization of A (cf. [2], Theorem T6, page 26): the process $\{A(t) - 2et,\ t \ge 0\}$ is a (right continuous) martingale with respect to $\{\mathscr{F}_t\}$, A is a pure jump process and at each of its jump points one component increases by 1 while the other stays constant.

The first step in our construction is to define a two-dimensional process $X = (X_1, X_2)$ on the integer lattice as follows. [Here and later, identities involving t are understood to hold almost surely (a.s.) for all $t \ge 0$ and $\int_0^t$ will mean $\int_{[0,t]}$.] Let

$$X_1(t) = A_1(t) - A_2(t), \tag{6.1}$$

$$\begin{aligned} X_2(t) = n &+ \int_0^t 1_{(-\infty,0)}(X_1(s-))\,dA_1(s) \\ &- \int_0^t 1_{(0,\infty)}(X_1(s-))\,dA_2(s). \end{aligned} \tag{6.2}$$

The integrands in (6.2) are predictable since $1_{(-\infty,0)}$ and $1_{(0,\infty)}$ can be written as limits of sequences of continuous functions and $X_1(\cdot -)$ is left continuous and adapted to $\{\mathscr{F}_t\}$. Thus, the integrals in (6.2) are well defined as stochastic integrals and define semimartingales with paths in D a.s. [A similar justification shows that the integrals in (6.13), (6.20)–(6.22), (6.33) and (6.34) below are well defined via stochastic calculus and yield adapted (to $\{\mathscr{F}_t\}$ or $\{\mathscr{G}_t\}$ as appropriate) processes with paths in D a.s.]

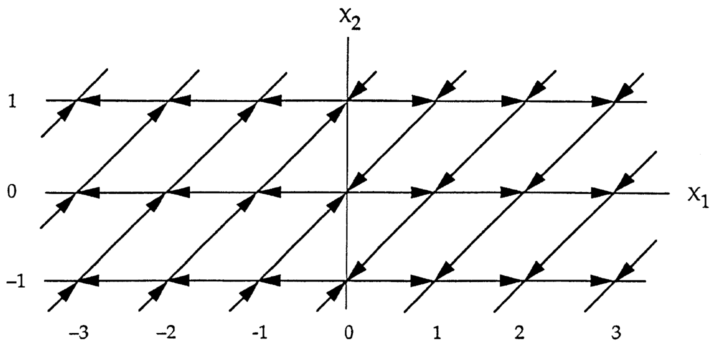

FIG. 4. *Transition structure of the Markov chain X.*

Equations (6.1) and (6.2) define a Markov chain X with $X(0) = (0, n)$ and the transition structure pictured in Figure 4. From each state (i, j) there are two possible transitions, and the possible directions differ depending on whether $i < 0$, $i = 0$ or $i > 0$.

The second step in our construction is to define another Markov chain Z that has $Z(0) = X(0)$ and the transition structure pictured in Figure 5 (again all transitions occur at rate 2). This will be accomplished by setting

$$Z_1 = X_1 \tag{6.3}$$

and defining Z_2 in terms of X_2 by a minor modification of the two-sided reflection mapping described in Section 2. To be specific, let

$$Z_2 = X_2 + Y_1 - Y_2, \tag{6.4}$$

where (Y_1, Y_2) is the least pair of nondecreasing, right continuous processes such that $Y_1(0) = Y_2(0) = 0$ and the process Z_2 defined by (6.4) satisfies

$$0 \le Z_2(t) \le n \quad \text{for all } t \ge 0. \tag{6.5}$$

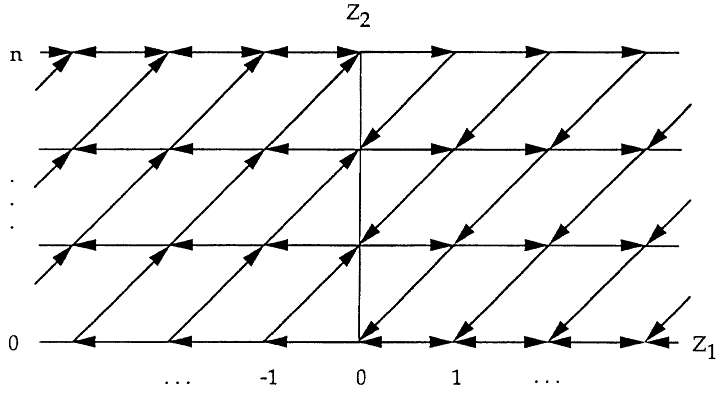

FIG. 5. *Transition structure of the reflected Markov chain Z (for n = 3).*

In words, Z_2 is obtained from X_2 by means of a reflection mapping that confines Z_2 to the interval $[0, n]$. The reflection mapping (η_1, η_2, ρ) described in Section 2 confines its image process to $[0, 1]$ and by rescaling by n we see that

$$Y_1 = n\eta_1(n^{-1}X_2), \qquad Y_2 = n\eta_2(n^{-1}X_2), \qquad Z_2 = n\rho(n^{-1}X_2).$$

The state space of Z is the strip Σ of integer lattice points pictured in Figure 5. In symbols,

$$\Sigma = \{(i, j): 0 \le j \le n\}. \tag{6.6}$$

[It is implicit here (and hereafter) that i and j are integers.] The transition structure of Z is the same as that for X except that the vertical component of any transition which would have carried Z above the upper boundary of Σ is "given back" (the horizontal component of the transition is still recorded), and similarly for transitions that would have carried Z below the lower boundary of Σ.

The third step in our construction is to modify Z by means of a time scale transformation, thus creating a new Markov chain U which is identical to Z except that time spent by Z in certain "forbidden" boundary states of Σ is eliminated. The forbidden states are those covered by the dark arrows in Figure 6, excluding the endpoints $(0, 0)$ and $(0, n)$. In symbols, the set of forbidden boundary states is

$$\Delta = \{(i, j): i > 0 \text{ and } j = 0, \text{ or } i < 0 \text{ and } j = n\} \tag{6.7}$$

and it will be convenient to define the complement

$$\Lambda = \Sigma - \Delta. \tag{6.8}$$

For reasons that will become apparent, the letters Δ and Λ may be considered mnemonic for "dead" and "live," respectively. We define continuous, nondecreasing processes δ and λ via

$$\delta(t) = \int_0^t 1_\Delta(Z(s))\, ds \tag{6.9}$$

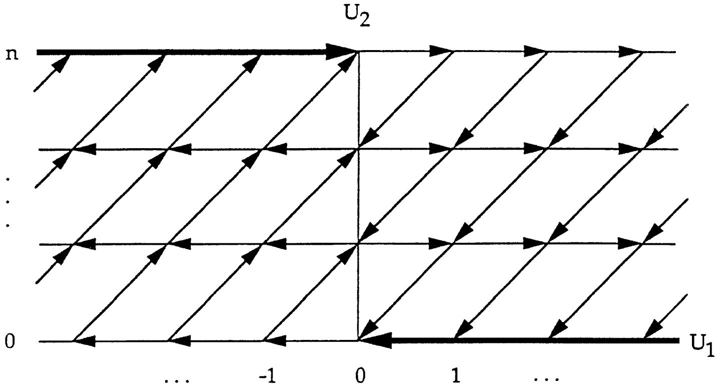

FIG. 6. *Transition structure of the Markov chain U (for $n = 3$).*

and

$$\lambda(t) = \int_0^t 1_\Lambda(Z(s))\, ds. \tag{6.10}$$

so that $\delta(t) + \lambda(t) = t$. Now let τ be the right continuous inverse of λ, meaning that

$$\tau(t) = \inf\{s \geq 0: \lambda(s) > t\}, \qquad t \geq 0. \tag{6.11}$$

To animate this definition, one may imagine a clock whose hands stop moving (in this sense they are dead) when Z is in Δ, but move at the normal rate (in this sense they are live) when Z is in Λ. Then $\tau(t)$ represents the amount of time required for the hands of this clock to advance by t time units. It follows from the nature of Z (in particular, Z_1 is a symmetric random walk) that a.s. $\tau(t) < \infty$ for each $t \geq 0$ and since $\lambda(t) \leq t$, we have that $\tau(t) \to \infty$ as $t \to \infty$. We now define

$$U(t) = Z(\tau(t)). \tag{6.12}$$

Because Z is Markov, a standard result (cf. [19], Section 65) on time change implies that U is Markov as well, and its transition structure is that pictured in Figure 6. That is, transitions of U are like those of Z except that, at the instant of a transition which would have caused entry into a forbidden state, there is immediate displacement to either $(0,0)$ or $(0,n)$, as shown by the dark arrows in Figure 6.

Lemma 6.2 below gives a decomposition of U_1 that will prove useful later. In preparation we define two-dimensional processes $\beta = (\beta_1, \beta_2)$ and $B = (B_1, B_2)$ via

$$\beta_j(t) = \int_0^t 1_\Lambda(Z(s-))\, dA_j(s), \qquad j = 1,2 \tag{6.13}$$

and

$$B(t) = \beta(\tau(t)). \tag{6.14}$$

Lemma 6.1. *Let $\mathscr{G}_t = \mathscr{F}_{\tau(t)}$ for $t \geq 0$. Then $\{B(t) - 2et, \mathscr{G}_t, t \geq 0\}$ is a martingale. Furthermore, B has the same distribution as A. That is, its components are independent Poisson processes starting from zero, each with arrival rate* 2.

Proof. Combining the definition (6.10) of λ with the definition (6.13) of β, we have that

$$\beta_j(t) - 2\lambda(t) = \int_0^t 1_\Lambda(Z(s-))\, d\big(A_j(s) - 2s\big). \tag{6.15}$$

Since Z is adapted to $\{\mathscr{F}_t\}$ and $\{A(t) - 2et, \mathscr{F}_t, t \geq 0\}$ is a martingale, it follows from stochastic calculus that $\{\beta(t) - 2e\lambda(t), \mathscr{F}_t, t \geq 0\}$ is a martingale. Then by the optional stopping theorem, for each integer $k \geq 0$, $\{\beta(\tau(t) \wedge k) - 2e\lambda(\tau(t) \wedge k), \mathscr{G}_t, t \geq 0\}$ is a martingale (cf. [5], Theorem 1.6). Now, for each fixed t, $\{\beta(\tau(t) \wedge k) - 2e\lambda(\tau(t) \wedge k), k \geq 0\}$ is uniformly integrable, since by quadratic variation estimates,

$$E\Big[|\beta(\tau(t) \wedge k) - 2e\lambda(\tau(t) \wedge k)|^2\Big] \leq 4E\big[\lambda(\tau(t) \wedge k)\big] \leq 4t.$$

It follows that $\{\beta(\tau(t)) - 2e\lambda(\tau(t)), \mathscr{G}_t, t \geq 0\}$ is a martingale (cf. [5], Proposition 1.8), but $\beta(\tau(t)) - 2e\lambda(\tau(t)) = B(t) - 2et$ by definition, and so the first statement in Lemma 6.1 is proved.

Obviously B is a pure jump process by construction and at each jump a single component increases by 1, which together with the aforementioned martingale property of $B(t) - 2et$ establishes the second statement (in Lemma 6.1 (cf. [2], page 25). □

LEMMA 6.2. *The Markov chain* $U = (U_1, U_2)$ *satisfies*

$$U_1(t) = (B_1(t) - B_2(t)) - J_1(t) + J_2(t), \tag{6.16}$$

where

(6.17) J_1 *and* J_2 *are right continuous, nondecreasing pure jump processes, with only finitely many jumps in each compact time interval,*

(6.18) $U(t) = (0,0)$ *for every t that is a jump time of* J_1

and

(6.19) $U(t) = (0,n)$ *for every t that is a jump time of* J_2.

REMARK. Because all processes here are right continuous, (6.18) says that all jumps of J_1 carry U into $(0,0)$ and similarly for (6.19). Combining this with our constructive definition of B, it is easy to show the following: at every time $t > 0$ when U enters state $(0,0)$ there is a unit increase in B_2 plus a possible jump of J_1; similarly, at every time $t > 0$ when U enters state $(0, n)$ there is unit increase in B_1 plus a possible jump in J_2.

PROOF OF LEMMA 6.2. At this point we need to distinguish notationally between the upper and lower sets of forbidden boundary states (see Figure 6). Let

$$\Phi = \{(i,j): i > 0 \text{ and } j = 0\} \quad \text{and} \quad \Psi = \{(i,j): i < 0 \text{ and } j = n\}.$$

Then $\Delta = \Phi \cup \Psi$, and Φ and Ψ are obviously disjoint. (Unfortunately, there is no mnemonic motivation for this choice of notation.) Now define

$$M(t) = \tfrac{1}{2}\int_0^t 1_\Lambda(Z(s-))\, dX_1(s), \tag{6.20}$$

$$N_1(t) = -\int_0^t 1_\Phi(Z(s-))\, dX_1(s), \tag{6.21}$$

$$N_2(t) = \int_0^t 1_\Psi(Z(s-))\, dX_1(s), \tag{6.22}$$

$$J_1(t) = N_1(\tau(t)) \quad \text{and} \quad J_2(t) = N_2(\tau(t)). \tag{6.23}$$

Because $X_1(t) = A_1(t) - A_2(t)$, it is immediate from (6.20), (6.13) and (6.14) that

$$2M(\tau(t)) = B_1(t) - B_2(t). \tag{6.24}$$

Also, because Λ, Φ and Ψ are disjoint and their union is the entire state space Σ of Z, we have from (6.20)–(6.22) that

$$X_1(t) = 2M(t) - N_1(t) + N_2(t).$$

Replacing t by $\tau(t)$ in the above and using (6.3), (6.12), (6.23) and (6.24), we arrive at (6.16).

It remains to show that the processes J_1 and J_2 defined by (6.23) satisfy (6.17)–(6.19). The proof is simplified by the fact that there are only finitely many jumps of X_1 in any finite time interval. We shall prove only the statements involving J_1 since those involving J_2 follow from symmetric arguments.

For (6.17), the right continuity of J_1 follows from that of N_1 and τ. For the proof of the jump property, let $T_1, T_2, \ldots$ denote the jump times of X_1, arranged in increasing order and let $T_0 = 0$. Then,

$$N_1(\tau(t)) = -\sum_{T_n \le \tau(t)} ((X_1(T_n) - X_1(T_{n-1})))1_\Phi(Z(T_{n-1})),$$

where $Z(T_{n-1}) \in \Phi$ implies $n > 1$ and $T_{n-1} \in [\tau(s-), \tau(s)) \neq \varnothing$ for some time s. Let $t_1, t_2, \ldots$ denote the (random) jump times of $\tau(\cdot)$, arranged in increasing order. Because of the nature of the Markov chain Z, almost surely there are only finitely many of these times in any compact time interval and $\tau(t_m) < \infty$ for each m. Note that $\tau(t_m -)$ and $\tau(t_m)$ must be jump times of X_1. Now the above expression for $N_1(\tau(t))$ may be rewritten as

$$\begin{aligned} N_1(\tau(t)) &= -\sum_{t_m \le t} \sum_{T_{n-1} \in [\tau(t_m -), \tau(t_m))} (X_1(T_n) - X_1(T_{n-1}))1_\Phi(Z(T_{n-1})) \\ &= -\sum_{t_m \le t} (X_1(\tau(t_m)) - X_1(\tau(t_m -)))1_\Phi(Z(\tau(t_m -))), \end{aligned} \tag{6.25}$$

where for $Z(\tau(t_m -)) \in \Phi$, $X_1(\tau(t_m)) = 0$ and $X_1(\tau(t_m -)) > 0$. It follows from this that J_1 is a nondecreasing pure jump process with only finitely many jumps in any compact time interval. Finally, to establish (6.18), let t be any jump time of J_1. Then by (6.25), $t = t_m$ for some m and $Z(\tau(t_m -)) \in \Phi$. Since $U(t) = Z(\tau(t)) \in \Lambda$ it follows that $\tau(t)$ is a time at which Z jumps from Φ to Λ, but this can only be true if $U(t) = Z(\tau(t)) = (0, 0)$ (see Figure 5). □

Let us denote by S the state space of V pictured in Figure 2 (it consists of integer lattice points lying within a parallelogram). Comparing Figures 2 and 5, we see that U has the same transition structure as the desired process V except that in constructing U we have allowed horizontal transitions that carry U outside the right and left boundaries of S. However, the cumulative effects of those transitions are corrected each time U reenters state $(0, 0)$ or $(0, n)$. Thus V_1 can be constructed from U_1 by simply collapsing the state space of the latter, as follows. Letting

$$L = \{(i, j): 0 \le j \le n \text{ and } i \le j - n\} \tag{6.26}$$

and

$$R = \{(i, j): 0 \le j \le n \text{ and } i \ge j\} \tag{6.27}$$

(the letters L and R are mnemonic for "left" and "right," respectively), we set

$$V_1(t) = \begin{cases} U_2(t) - n, & \text{if } U(t) \in L, \\ U_2(t), & \text{if } U(t) \in R, \\ U_1(t), & \text{otherwise,} \end{cases} \tag{6.28}$$

$$V_2(t) = U_2(t). \tag{6.29}$$

To repeat, the process V defined by (6.28) and (6.29) has the desired transition structure pictured in Figure 2 and thus it provides a Markovian representation of the closed queueing network under consideration. Server 1 is idle if and only if V is on the right boundary of its state space S and, in like fashion, cumulative idleness of server 2 is equivalent to cumulative occupation time of the left boundary. From (6.28) we see that $V \in R$ if and only if $U \in R$ and, similarly, $V \in L$ if and only if $U \in L$, so the cumulative idleness processes (see Section 2) can be written as

$$I_1(t) = \int_0^t 1_R(V(s))\, ds = \int_0^t 1_R(U(s))\, ds, \tag{6.30}$$

$$I_2(t) = \int_0^t 1_L(V(s))\, ds = \int_0^t 1_L(U(s))\, ds. \tag{6.31}$$

As an alternative to (6.28), one can describe the construction of V_1 from U as follows: each time there occurs a horizontal transition of U that begins from a state in $L \setminus \{(0, n)\}$ or $R \setminus \{(0, 0)\}$ (such transitions always have the effect of carrying U farther from the desired state space S), the transition is simply "given back." That is, one can reexpress (6.28) in terms of the processes B_j that occur in (6.16) by writing

$$V_1(t) = (B_1(t) - B_2(t)) - K_1(t) + K_2(t), \tag{6.32}$$

where

$$K_1(t) = \int_0^t 1_R(U(s-))\, dB_1(s) \tag{6.33}$$

and

$$K_2(t) = \int_0^t 1_L(U(s-))\, dB_2(s). \tag{6.34}$$

The processes $(K_1 - J_1)$ and $(K_2 - J_2)$ are both nonnegative and they cannot both be strictly positive at the same time, and when U is at $(0, 0)$ or $(0, n)$, $K_1 - J_1$ and $K_2 - J_2$ are both zero. Thus, from (6.16) and (6.32) it follows that

$$|U_1(t) - V_1(t)| = |K_1(t) - J_1(t)| + |K_2(t) - J_2(t)|. \tag{6.35}$$

A key step in the proof of Theorem 5.1 is to show that $(K - J)$ vanishes under our heavy traffic scaling and hence V is indistinguishable from U in the heavy traffic limit. Moreover, the parallel structure of (6.30), (6.31) and (6.33), (6.34) will allow us to show that the processes K and $2I$ are asymptotically indistinguishable under heavy traffic scaling, implying asymptotic equivalence of J and $2I$ under that scaling.

7. Construction of the limit process. In this section we rigorously construct the limit process that was described heuristically in Section 5 and we prove a semimartingale decomposition for this process. Here the term diffusion will mean a continuous strong Markov process.

Let X_1^* be a one-dimensional Brownian motion with drift parameter equal to zero, variance parameter equal to 4 and such that $X_1^*(0) = 0$. Define

$$X_2^*(t) = 1 + 2\int_0^t 1_{(-\infty,0)}(X_1^*(s))\,ds - 2\int_0^t 1_{(0,\infty)}(X_1^*(s))\,ds. \tag{7.1}$$

Then for any (possibly random) time $T \geq 0$,

$$\begin{aligned} X_2^*(t+T) = X_2^*(T) &+ 2\int_0^t 1_{(-\infty,0)}(X_1^*(s+T))\,ds \\ &- 2\int_0^t 1_{(0,\infty)}(X_1^*(s+T))\,ds. \end{aligned} \tag{7.2}$$

It follows from this and the strong Markov property of X_1^* that the two-dimensional process $X^* = (X_1^*, X_2^*)$ is a diffusion process.

We now construct a reflected diffusion process Z^* that lives in the strip $S^* \equiv \mathbb{R} \times [0,1]$ and that has normal reflection at the boundary of S^*. This is achieved by applying the two-sided reflection mapping (η_1, η_2, ρ) described in Section 2 to X_2^*. We define $Z^* = (Z_1^*, Z_2^*)$:

$$Z_1^* = X_1^*, \tag{7.3}$$

$$Z_2^* = \rho(X_2^*) = X_2^* + Y_1^* - Y_2^*, \tag{7.4}$$

where $Y_1^* = \eta_1(X_2^*)$ and $Y_2^* = \eta_2(X_2^*)$. It follows from the uniqueness cited in Proposition 2.2 that for any (possibly random) time $T \geq 0$ and for all $t \geq 0$,

$$\begin{aligned} Z_2^*(t+T) &= \rho(Z_2^*(T) + X_2^*(t+T) - X_2^*(T)), \\ Y_1^*(t+T) - Y_1^*(T) &= \eta_1(Z_2^*(T) + X_2^*(t+T) - X_2^*(T)), \\ Y_2^*(t+T) - Y_2^*(T) &= \eta_2(Z_2^*(T) + X_2^*(t+T) - X_2^*(T)). \end{aligned} \tag{7.5}$$

This, together with (7.2) and the stationarity and independence of the increments of $X_1^* = Z_1^*$, implies that $Z^* = (Z_1^*, Z_2^*)$ is a diffusion process.

One can heuristically describe the behavior of Z^* as follows (see Figure 7). Of course, Z_1^* is a one-dimensional Brownian motion with zero drift and variance parameter 4. When Z^* is in the interior of S^*, Z_2^* is a drift process where its state dependent drift is $+2$ if $Z_1^* < 0$ and -2 if $Z_1^* > 0$. (The drift of Z_2^* when $Z_1^* = 0$ does not have to be carefully specified because the amount of time that Z_1^* is zero has zero Lebesgue measure almost surely.) At the boundaries of the strip, Z^* is confined to S^* by instantaneous reflection (or pushing) where the directions of reflection are vertical and up or down as $Z_2^* = 0$ or $Z_2^* = 1$, respectively.

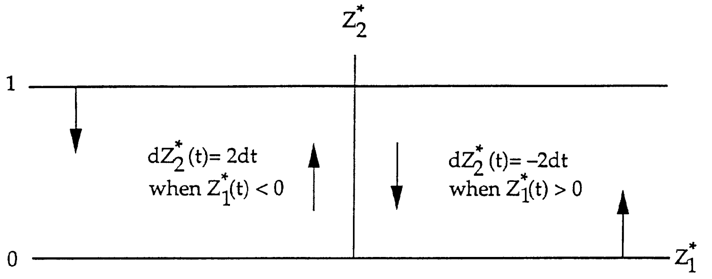
Z_2^*
1
0
Z_1^*
$dZ_2^*(t) = 2dt$ when $Z_1^*(t) < 0$
$dZ_2^*(t) = -2dt$ when $Z_1^*(t) > 0$

FIG. 7. *Drifts and directions of reflection for the diffusion Z^*.*

Now define

$$\Delta^* = \{z^* \in S^*: z_1^* > 0 \text{ and } z_2^* = 0, \text{ or } z_1^* < 0 \text{ and } z_2^* = 1\}, \tag{7.6}$$

$$\Lambda^* = S^* \setminus \Delta^*, \tag{7.7}$$

$$\lambda^*(t) = \int_0^t 1_{\Lambda^*}(Z^*(s))\, ds \tag{7.8}$$

and let τ^* be the right continuous inverse of λ^* defined by

$$\tau^*(t) = \inf\{s \ge 0: \lambda^*(s) > t\}. \tag{7.9}$$

By analogy with the notation in Section 6, we decompose Δ^* into lower and upper boundary parts:

$$\Phi^* = \{z^* \in S^*: z_1^* > 0 \text{ and } z_2^* = 0\} \quad \text{and}$$
$$\Psi^* = \{z^* \in S^*: z_1^* < 0 \text{ and } z_2^* = 1\},$$

so that $\Delta^* = \Phi^* \cup \Psi^*$.

The following lemma is proved in Appendix A.

LEMMA 7.1. *Almost surely*:

(i) $\lambda^*(\infty) \equiv \lim_{t\to\infty} \lambda^*(t) = \infty$ *and hence* $\tau^*(t) < \infty$ *for all* $t \ge 0$;
(ii) $\lambda^*(t) > 0$ *for all* $t > 0$ *and hence* $\tau^*(0) = 0$.

Define

$$U^*(t) = Z^*(\tau^*(t)). \tag{7.10}$$

Now once Z^* hits Δ^*, it remains in that set until it reaches one of the endpoints $(0, 0)$ or $(0, 1)$. Moreover, the time change τ^* deletes the time that Z^* is in Δ^*. It then follows from an easy argument by contradiction that U^* lives in Λ^*. We now obtain a semimartingale decomposition of U^*, which in particular yields (5.6). The process

$$M^*(t) = \tfrac{1}{2}\int_0^t 1_{\Lambda^*}(Z^*(s))\, dX_1^*(s)$$

is a continuous local martingale with respect to the filtration generated by Z^* and its quadratic variation pocess is given by

$$[M^*](t) = \int_0^t 1_{A^*}(Z^*(s))\,ds = \lambda^*(t).$$

It is well known that any continuous local martingale M^* with $[M^*](\infty) = \infty$ a.s. can be time changed to a Brownian motion (cf. Theorem 9.3 of [5]). Thus,

$$W^*(t) \equiv M^*(\tau^*(t)) = \tfrac{1}{2}\int_0^{\tau^*(t)} 1_{\Lambda^*}(Z^*(s))\,dX_1^*(s) \tag{7.11}$$

is a driftless Brownian motion with variance parameter equal to 1.

Combining the above relations we have a.s. for all $t \geq 0$,

$$\begin{aligned} U_1^*(t) &= 2W^*(t) + \int_0^{\tau^*(t)} 1_{\Delta^*}(Z^*(s))\,dX_1^*(s) \\ &= 2W^*(t) - J_1^*(t) + J_2^*(t), \end{aligned} \tag{7.12}$$

where

$$\begin{aligned} J_1^*(t) &\equiv N_1^*(\tau^*(t)) = -\int_0^{\tau^*(t)} 1_{\Phi^*}(Z^*(s))\,dX_1^*(s), \\ J_2^*(t) &\equiv N_2^*(\tau^*(t)) = \int_0^{\tau^*(t)} 1_{\Psi^*}(Z^*(s))\,dX_1^*(s), \\ N_1^*(t) &= -\int_0^t 1_{\Phi^*}(Z^*(s))\,dX_1^*(s), \\ N_2^*(t) &= \int_0^t 1_{\Psi^*}(Z^*(s))\,dX_1^*(s). \end{aligned} \tag{7.13}$$

The following lemma is proved in Appendix A using an approximate decomposition of the J_j^* according to excursions of Z^* from (0, 0) and (0, 1). With this, the justification of the decomposition (5.6) is complete.

LEMMA 7.2. *Almost surely*:

(i) J_1^* *and* J_2^* *are nondecreasing*;
(ii) $J_1^*(0) = J_2^*(0) = 0$;
(iii) J_1^* *can have a point of increase at time t only if* $U^*(t) = (0,0)$ *and* J_2^* *can have a point of increase at time t only if* $U^*(t) = (0,1)$.

To obtain the decomposition (5.7) of U_2^*, we first need to establish that the supports of Y_1^* and Y_2^* as integrators are contained in $\{t \geq 0\colon Z^*(t) \in \Delta^*\}$. For this, let

$$\begin{aligned} \Lambda_1^* &= \{z^* \in S^*\colon z_1^* \leq 0 \text{ and } z_2^* = 0\}, \\ \Lambda_2^* &= \{z^* \in S^*\colon z_1^* \geq 0 \text{ and } z_2^* = 1\}. \end{aligned}$$

The following lemma is proved in Appendix A using estimates obtained by applying Itô's formula to suitable test functions.

LEMMA 7.3. *We have*

$$\int_0^\infty 1_{\Lambda_1^* \cup \Lambda_2^*}(Z^*(s))\,ds = 0 \quad a.s., \tag{7.14}$$

$$\int_0^\infty 1_{\Lambda_1^*}(Z^*(s))\,dY_1^*(s) = 0 \quad a.s., \tag{7.15}$$

$$\int_0^\infty 1_{\Lambda_2^*}(Z^*(s))\,dY_2^*(s) = 0 \quad a.s. \tag{7.16}$$

REMARK. Now, by definition, Y_1^* can increase only when $Z_2^* = 0$. Combining this with (7.15), we see that as an integrator Y_1^* only charges the set of times for which Z^* is in $\Phi^* \subset \Delta^*$. Similarly, Y_2^* only charges the set of times for which Z^* is in Ψ^*. Furthermore, it follows from (7.14) that since X_2^* is a drift process [cf. (7.1)], X_2^* as an integrator can only charge the set of times for which Z_2^* is in $(0,1)$ or Z^* is in Δ^*.

Combining the above results with (7.4), we have

$$Z_2^*(t) = 1 + \int_0^t 1_{(0,1)}(Z_2^*(s))\,dX_2^*(s) + \int_0^t 1_{\Delta^*}(Z^*(s))\,dZ_2^*(s). \tag{7.17}$$

The following lemma is shown in Appendix A.

LEMMA 7.4. *Almost surely, for all $t \ge 0$,*

$$\int_0^t 1_{\Delta^*}(Z^*(s))\,dZ_2^*(s) = 0 \tag{7.18}$$

and

$$\begin{aligned} Z_2^*(t) &= 1 + 2\int_0^t 1_{(-\infty,0)\times(0,1)}(Z^*(s))\,ds \\ &\quad - 2\int_0^t 1_{(0,\infty)\times(0,1)}(Z^*(s))\,ds. \end{aligned} \tag{7.19}$$

REMARK. Lemma 7.4 corresponds to the intuitive fact that when Z^* is in Δ^*, it does not move vertically.

Since $d\lambda^*(s) = ds$ on $\{s \ge 0\colon Z_2^*(s) \in (0,1)\}$, we can change variables $[s = \tau^*(u)]$ in the integrations in (7.19) and use (7.10) to obtain

$$\begin{aligned} U_2^*(t) &= 1 + 2\int_0^t 1_{(-\infty,0)\times(0,1)}(U^*(u))\,du \\ &\quad - 2\int_0^t 1_{(0,\infty)\times(0,1)}(U^*(u))\,du. \end{aligned} \tag{7.20}$$

Finally, note that by the reverse change of variables and Lemma 7.3,

$$
\begin{aligned}
&\int_0^t 1_{(0,1)}(U_2^*(u))\,du \\
&\qquad = \int_0^{\tau^*(t)} 1_{(0,1)}(Z_2^*(s))\,ds \\
&\qquad = \tau^*(t) - \int_0^{\tau^*(t)} 1_{\{0,1\}}(Z_2^*(s))\,ds \\
&\qquad = \tau^*(t) - \int_0^{\tau^*(t)} 1_{\Delta^*}(Z^*(s))\,ds = \lambda^*(\tau^*(t)) = t.
\end{aligned}
\tag{7.21}
$$

Hence (5.7) holds.

8. Proof of the heavy traffic limit theorem. In this section we prove Theorem 5.1. To elucidate the skeleton of this argument, we defer intricate proofs to Appendix B. In the statements of convergence in this section and Appendix B, we shall frequently suppress the qualifier "$n \to \infty$" when its implicit presence is clear from the context.

From (3.3)–(3.6) and (6.30), (6.31), we see that the queue-length and idleness processes for the nth system can be represented in terms of the Markov chain V, which in turn can be constructed from the process X defined in Section 6. As in Section 5, we shall use a superscript n to indicate the dependence on n of the processes defined in Section 6. Thus, X will be written as X^n, but the process A will not have a superscript n because it does not vary with n. We begin by rescaling X^n so that its first component has a CLT type of rescaling like that for $\hat{Q}_2^n, \hat{Q}_4^n$ and its second component has a law of large numbers type of rescaling like that for $\hat{Q}_1^n, \hat{Q}_3^n$ [cf. (5.2) and (5.3)]:

$$
\hat{A}_j^n(t) \equiv \frac{1}{\sqrt{n}}\bigl(A_j(nt) - 2nt\bigr), \qquad j = 1, 2, \tag{8.1}
$$

$$
\hat{X}_1^n(t) \equiv \frac{1}{\sqrt{n}} X_1^n(nt) = \hat{A}_1^n(t) - \hat{A}_2^n(t), \tag{8.2}
$$

$$
\bar{A}_j^n(t) \equiv \frac{1}{n} A_j(nt), \qquad j = 1, 2, \tag{8.3}
$$

$$
\begin{aligned}
\hat{X}_2^n(t) &\equiv \frac{1}{n} X_2^n(nt) \\
&= 1 + \int_0^t 1_{(-\infty,0)}\bigl(\hat{X}_1^n(s-)\bigr)\,d\bar{A}_1^n(s) \\
&\qquad - \int_0^t 1_{(0,\infty)}\bigl(\hat{X}_1^n(s-)\bigr)\,d\bar{A}_2^n(s).
\end{aligned}
\tag{8.4}
$$

Similarly,

$$\hat{Z}_1^n(t) \equiv \frac{1}{\sqrt{n}} Z_1^n(nt) = \hat{X}_1^n(t), \tag{8.5}$$

$$\hat{Z}_2^n(t) \equiv \frac{1}{n} Z_2^n(nt) = \hat{X}_2^n(t) + \hat{Y}_1^n(t) - \hat{Y}_2^n(t), \tag{8.6}$$

where

$$\hat{Y}_j^n(t) \equiv \frac{1}{n} Y_j^n(nt) = \eta_j\big(\hat{X}_2^n\big)(t), \qquad j = 1, 2. \tag{8.7}$$

Let

$$A^*(t) = 2t \tag{8.8}$$

and let B_j^*, $j = 1, 2$, be two independent driftless one-dimensional Brownian motions such that each starts from the origin and has variance parameter equal to 2. Without loss of generality, we suppose that $X_1^* = B_1^* - B_2^*$. Now, by the functional law of large numbers and central limit theorems for the independent Poisson processes A_j, $j = 1, 2$ (cf. [1], Theorem 17.3),

$$\bar{A}_j^n \Rightarrow A^* \quad \text{and} \quad \hat{A}_j^n \Rightarrow B_j^* \quad \text{in the } \mathbf{J}_1 \text{ topology, for } j = 1, 2.$$

Indeed, since A^* is deterministic and A_1 (respectively, B_1^*) is independent of A_2 (respectively, B_2^*), we have (cf. [1], page 27)

$$\left(\bar{A}_1^n, \hat{A}_1^n, \bar{A}_2^n, \hat{A}_2^n\right) \Rightarrow (A^*, B_1^*, A^*, B_2^*)$$

with the product topology on $D^4 = D \times D \times D \times D$, where each copy of D is endowed with the $\mathbf{J}_1$ topology. Finally, since $\hat{X}_1^n$ is the difference of $\hat{A}_1^n$ and $\hat{A}_2^n$, and the limit processes B_1^*, B_2^* are continuous, it follows [cf. Proposition 2.1(ii)] that

$$\left(\hat{X}_1^n, \bar{A}_1^n, \hat{A}_1^n, \bar{A}_2^n, \hat{A}_2^n\right) \Rightarrow (X_1^*, A^*, B_1^*, A^*, B_2^*), \tag{8.9}$$

with the product topology on D^5, where each copy of D has the $\mathbf{J}_1$ topology. In fact, this weak convergence holds with the $\mathbf{J}_1$ topology on D^5 since the limit processes are all continuous. To see this, note that by the Skorokhod representation theorem ([8], Theorem 3.1.8), we could suppose that the convergence in (8.9) is a.s. in the product topology and since the limit processes are continuous, this convergence is almost surely u.o.c. for each component [cf. Proposition 2.1(i)] and hence u.o.c. for the vector of components, which implies convergence in the $\mathbf{J}_1$ topology on D^5.

Observe that by (8.4), $\hat{X}_2^n$ is defined from $\hat{X}_1^n$ and the $\bar{A}_j^n$, $j = 1, 2$, and by (7.1), X_2^* is defined from X_1^*. Using these representations and the weak convergence in (8.9), together with (B.4) to take care of the discontinuity of the integrands in (7.1), the following lemma is proved in Appendix B.

LEMMA 8.1. *We have*

$$\begin{aligned}(8.10)\qquad &\left(\hat X_1^n, \hat X_2^n, \bar A_1^n, \hat A_1^n, \bar A_2^n, \hat A_2^n\right)\\ &\qquad\Rightarrow (X_1^*, X_2^*, A^*, B_1^*, A^*, B_2^*) \text{ in the } \mathbf{J}_1 \text{ topology}.\end{aligned}$$

Now by Proposition 2.3 and the continuous mapping theorem, we can add $\hat Z_2^n = \rho(\hat X_2^n) \Rightarrow \rho(X_2^*) = Z_2^*$ as an additional component in the convergence in (8.10). Also, $\hat Z_1^n = \hat X_1^n$ and $Z_1^* = X_1^*$. Hence we have

$$\begin{aligned}(8.11)\qquad &\left(\hat Z_1^n, \hat Z_2^n, \hat X_1^n, \hat X_2^n, \bar A_1^n, \hat A_1^n, \bar A_2^n, \hat A_2^n\right)\\ &\qquad\Rightarrow (Z_1^*, Z_2^*, X_1^*, X_2^*, A^*, B_1^*, A^*, B_2^*)\end{aligned}$$

in the $\mathbf{J}_1$ topology.

We now continue with the normalization of quantities introduced in Section 6 to define $\hat\Lambda^n$, $\hat\lambda^n$, $\hat\tau^n$ and $\hat U^n$. Let

$$\begin{aligned}(8.12)\qquad \hat\Lambda^n \equiv \Bigg\{\left(\frac{i}{\sqrt n}, \frac{j}{n}\right)&: 0 < j < n, \text{ or } i \le 0 \text{ and } j = 0,\\ &\text{or } i \ge 0 \text{ and } j = n\Bigg\},\end{aligned}$$

$$(8.13)\qquad \hat\lambda^n(t) \equiv \frac{1}{n}\lambda^n(nt) = \int_0^t 1_{\hat\Lambda^n}\big(\hat Z^n(s)\big)\,ds = \int_0^t 1_{\Lambda^*}\big(\hat Z^n(s)\big)\,ds,$$

$$(8.14)\qquad \hat\phi^n(t) \equiv \int_0^t 1_{\Phi^*}\big(\hat Z^n(s)\big)\,ds, \qquad \hat\psi^n(t) \equiv \int_0^t 1_{\Psi^*}\big(\hat Z^n(s)\big)\,ds,$$

$$(8.15)\qquad \hat\tau^n(t) \equiv \frac{1}{n}\tau^n(nt) = \inf\{s \ge 0: \hat\lambda^n(s) > t\},$$

$$(8.16)\qquad \hat U_1^n(t) \equiv \frac{1}{\sqrt n} U_1^n(nt),$$

$$(8.17)\qquad \hat U_2^n(t) \equiv \frac{1}{n} U_2^n(nt),$$

so that

$$(8.18)\qquad \hat U^n(t) = \hat Z^n(\hat\tau^n(t)).$$

Now, by (8.16), (6.16) and (6.20)–(6.24),

$$(8.19)\qquad \hat U_1^n(t) = 2\hat W^n(t) - \hat J_1^n(t) + \hat J_2^n(t),$$

where

$$\begin{aligned}(8.20)\qquad &\hat W^n(t) = \hat M^n(\hat\tau^n(t)),\\ &\hat J_j^n(t) \equiv \frac{1}{\sqrt n} J_j^n(nt) = \hat N_j^n(\hat\tau^n(t)), \qquad j = 1, 2,\end{aligned}$$

$$(8.21)\qquad \hat M^n(t) \equiv \frac{1}{\sqrt n} M^n(nt) = \frac{1}{2}\int_0^t 1_{\Lambda^*}\big(\hat Z^n(s-)\big)\,d\hat Z_1^n(s),$$

$$\hat{N}_1^n(t) \equiv \frac{1}{\sqrt{n}} N_1^n(nt) = -\int_0^t 1_{\Phi^*}\bigl(\hat{Z}^n(s-)\bigr)\, d\hat{Z}_1^n(s), \tag{8.22}$$

$$\hat{N}_2^n(t) \equiv \frac{1}{\sqrt{n}} N_2^n(nt) = \int_0^t 1_{\Psi^*}\bigl(\hat{Z}^n(s-)\bigr)\, d\hat{Z}_1^n(s). \tag{8.23}$$

In addition to the definitions of starred processes made in Section 7, we define

$$\phi^*(t) = \int_0^t 1_{\Phi^*}(Z^*(s))\, ds, \qquad \psi^*(t) = \int_0^t 1_{\Psi^*}(Z^*(s))\, ds. \tag{8.24}$$

The following lemma is proved in Appendix B.

LEMMA 8.2. *We have*

$$\begin{aligned}&\left(\hat{Z}_1^n, \hat{Z}_2^n, \hat{X}_1^n, \hat{X}_2^n, \bar{A}_1^n, \hat{A}_1^n, \bar{A}_2^n, \hat{A}_2^n, \hat{\lambda}^n, \hat{\phi}^n, \hat{\psi}^n, \hat{M}^n, \hat{N}_1^n, \hat{N}_2^n\right)\\ &\qquad \Rightarrow (Z_1^*, Z_2^*, X_1^*, X_2^*, A^*, B_1^*, A^*, B_2^*, \lambda^*, \phi^*, \psi^*, M^*, N_1^*, N_2^*)\end{aligned} \tag{8.25}$$

in the $\mathbf{J}_1$ *topology.*

Having established the above preliminaries, we now turn to the main part of the proof of Theorem 5.1. By Skorokhod's representation theorem we may assume that the convergence in (8.25) is almost surely u.o.c. as $n \to \infty$. Now, $\hat{\lambda}^n \to \lambda^*$ u.o.c. almost surely implies (cf. [13], page 1018) that a.s. $\hat{\tau}^n(t) \to \tau^*(t)$ as $n \to \infty$ at each continuity point t of τ^*. Then, since a.s. $\hat{M}^n \to M^*$ u.o.c. as $n \to \infty$ and M^* is constant on any interval where its quadratic variation process λ^* is constant (cf. [5], page 189), it follows from Lemma 2.3 of Kurtz [13] that a.s. as $n \to \infty$,

$$\hat{W}^n \equiv \hat{M}^n(\hat{\tau}^n) \to M^*(\tau^*) \equiv W^* \quad \text{u.o.c.} \tag{8.26}$$

Furthermore, for $j = 1, 2$, since N_j^* is continuous, by the proof of Lemma 2.3(a) of Kurtz [13], a.s. as $n \to \infty$,

$$\hat{J}_j^n(t) \equiv \hat{N}_j^n(\hat{\tau}^n(t)) \to N_j^*(\tau^*(t)) \equiv J_j^*(t) \tag{8.27}$$

at all continuity points t of τ^*. By the right continuity of τ^*, $t = 0$ is a continuity point of τ^*. Then, since $\hat{J}_j^n$ and J_j^* are a.s. nonnegative and nondecreasing (cf. Lemmas 6.2 and 7.2), it follows from Proposition 2.1(iii) that a.s. as $n \to \infty$,

$$\hat{J}_j^n \to J_j^* \quad \text{in the } \mathbf{M}_1 \text{ topology as } n \to \infty \text{ for } j = 1, 2. \tag{8.28}$$

By Lemma 7.2, a.s. the sets of times of discontinuity (jumps) of J_1^* and J_2^* are disjoint. Then since $\mathbf{J}_1$ convergence implies $\mathbf{M}_1$ convergence, it follows from (8.26), (8.28), (8.19), (7.12) and Proposition 2.1(ii) that a.s. as $n \to \infty$,

$$\hat{U}_1^n \to U_1^* \quad \text{in the } \mathbf{M}_1 \text{ topology.} \tag{8.29}$$

Another application of Lemma 2.3 of Kurtz [13], together with the a.s. convergence of $\{(\hat{Z}_2^n, \hat{\lambda}^n)\}_{n=1}^{\infty}$ to (Z_2^*, λ^*) u.o.c. and the fact (7.18) that a.s. Z_2^* remains constant on any interval where λ^* is constant, yields a.s. as $n \to \infty$,

$$\hat{U}_2^n = \hat{Z}_2^n(\hat{\tau}^n) \to Z_2^*(\tau^*) = U_2^* \quad \text{u.o.c.} \tag{8.30}$$

Rescaling V^n in the same way as U^n, we define

$$\hat{V}_1^n(t) = \frac{1}{\sqrt{n}} V_1^n(nt), \qquad \hat{V}_2^n(t) = \frac{1}{n} V_2^n(nt), \tag{8.31}$$

so that by (6.28), (6.29), (8.16) and (8.17), we have

$$\hat{V}_1^n(t) = \begin{cases} \sqrt{n}\,\hat{U}_2^n(t) - \sqrt{n}, & \text{if } \hat{U}^n(t) \in \hat{L}^n, \\ \sqrt{n}\,\hat{U}_2^n(t), & \text{if } \hat{U}^n(t) \in \hat{R}^n, \\ \hat{U}_1^n(t), & \text{otherwise,} \end{cases} \tag{8.32}$$

$$\hat{V}_2^n(t) = \hat{U}_2^n(t), \tag{8.33}$$

where

$$\hat{L}^n = \left\{ \left(\frac{i}{\sqrt{n}}, \frac{j}{n} \right) : 0 \le \frac{j}{n} \le 1 \text{ and } \frac{i}{\sqrt{n}} \le \sqrt{n}\left(\frac{j}{n} \right) - \sqrt{n} \right\}, \tag{8.34}$$

$$\hat{R}^n = \left\{ \left(\frac{i}{\sqrt{n}}, \frac{j}{n} \right) : 0 \le \frac{j}{n} \le 1 \text{ and } \frac{i}{\sqrt{n}} \ge \sqrt{n}\left(\frac{j}{n} \right) \right\}. \tag{8.35}$$

Then the state space of $\hat{V}^n$ is

$$\begin{aligned} \hat{S}^n \equiv \Bigg\{ \left(\frac{i}{\sqrt{n}}, \frac{j}{n} \right) &: 0 \le \frac{j}{n} \le 1 \text{ and} \\ &\sqrt{n}\left(\frac{j}{n} \right) - \sqrt{n} \le \frac{i}{\sqrt{n}} \le \sqrt{n}\left(\frac{j}{n} \right) \Bigg\}. \end{aligned} \tag{8.36}$$

Using (6.24), (6.20), (6.3), (8.5), (8.15)–(8.17), (8.20), (8.21), (6.26), (6.27), (8.34) and (8.35) to follow the rescaling (8.31) of the representation (6.32)–(6.34) for V_1^n, we obtain

$$\hat{V}_1^n(t) = 2\hat{W}^n(t) - \hat{K}_1^n(t) + \hat{K}_2^n(t), \tag{8.37}$$

where

$$\hat{K}_1^n(t) \equiv \frac{1}{\sqrt{n}} K_1^n(nt) = \int_0^t 1_{\hat{R}^n}\big(\hat{U}^n(s-)\big)\, d\hat{B}_1^n(s), \tag{8.38}$$

$$\hat{K}_2^n(t) \equiv \frac{1}{\sqrt{n}} K_2^n(nt) = \int_0^t 1_{\hat{L}^n}\big(\hat{U}^n(s-)\big)\, d\hat{B}_2^n(s), \tag{8.39}$$

$$\hat{B}_j^n(t) = \frac{1}{\sqrt{n}} B_j^n(nt), \qquad j = 1, 2. \tag{8.40}$$

Note that $\hat{K}_1^n$ is the magnitude of the cumulative excess movement to the right (in $\hat{R}^n$) associated with the movement of $\hat{U}^n$ outside of $\hat{S}^n$. Similarly, $\hat{K}_2^n$ is the magnitude of the excess movement of $\hat{U}^n$ to the left (in $\hat{L}^n$) outside of $\hat{S}^n$. These movements need to be "given back" in order to recover $\hat{V}^n$ from $\hat{U}^n$. From (6.35) we have

$$(8.41) \qquad \left|\hat{U}_1^n(t) - \hat{V}_1^n(t)\right| = \left|\hat{K}_1^n(t) - \hat{J}_1^n(t)\right| + \left|\hat{K}_2^n(t) - \hat{J}_2^n(t)\right|.$$

By Lemma 7.3 and the construction of U^* from Z^* by deletion of time, almost surely U^* spends zero Lebesgue time in $\Lambda_1^* \cup \Lambda_2^*$ and hence is in the interior $S^\circ = \Lambda^* \setminus (\Lambda_1^* \cup \Lambda_2^*)$ of the strip S^* for all but a set of times of Lebesgue measure zero.

Let Ω denote the sample space on which all of our processes are defined. When we wish to indicate the dependence of a given process A on $\omega \in \Omega$, we shall use $A(\cdot, \omega)$ to denote the sample path of the process associated with ω. In particular, $A(t, \omega)$ will denote the position of that path at time t. When there is no need to indicate the dependence of the process on ω, we shall simply write $A(t)$ for the (random) value of the process A at time t. Let $\omega \in \Omega$ be such that the following hold:

1. $U^*(\cdot, \omega)$ is in S° for all but a set of times of Lebesgue measure zero.
2. $\hat{K}_j^n(\cdot, \omega)$ is nondecreasing and $\hat{K}_j^n(0, \omega) = 0$ for all n and $j = 1,2$.
3. The properties of J_j^*, $j = 1,2$, listed in Lemma 7.2 hold at ω.
4. $Z^*(\cdot, \omega)$ is continuous.
5. As $n \to \infty$,

$$(8.42) \qquad \left(\hat{W}^n, \hat{U}^n, \hat{J}_1^n, \hat{J}_2^n\right)(\cdot, \omega) \to (W^*, U^*, J_1^*, J_2^*)(\cdot, \omega)$$

with the product topology on D^5, where each copy of D has the $\mathbf{M}_1$ topology.

The set of such ω has probability 1 [cf. (6.33), (6.34), Lemma 7.2, (8.26), (8.28) and (8.29)]. Now, if $U^*(t, \omega) \in S^\circ$, then t is a continuity point of $\tau^*(\omega)$ and hence of $U^*(\cdot, \omega)$, and so by the $\mathbf{M}_1$ convergence,

$$(8.43) \qquad \hat{U}^n(t, \omega) \to U^*(t, \omega) \quad \text{as } n \to \infty.$$

Let

$$(8.44) \qquad \bar{S}^n = \left\{(x, y) \colon 0 \le y \le 1 \text{ and } \sqrt{n}\, y - \sqrt{n} \le x \le \sqrt{n}\, y\right\}.$$

Note that $\hat{S}^n \subset \bar{S}^n$ for all n, and the interior $(\bar{S}^n)^\circ$ of $\bar{S}^n$ is increasing with n and the union over n of these interiors equals S°. It follows that there is $n_1 = n_1(t, \omega)$ such that $U^*(t, \omega) \in (\bar{S}^n)^\circ$ for all $n \ge n_1$, and then by (8.43) there is $n_2 = n_2(t, \omega) \ge n_1(t, \omega)$ such that $\hat{U}^n(t, \omega) \in (\bar{S}^{n_2})^\circ \subset (\bar{S}^n)^\circ$ for all $n \ge n_2$. It follows [cf. (8.32)] that $\hat{V}_1^n(t, \omega) = \hat{U}_1^n(t, \omega)$ for all $n \ge n_2$. Then by (8.41),

$$(8.45) \qquad \hat{J}_j^n(t, \omega) = \hat{K}_j^n(t, \omega) \quad \text{for all } n \ge n_2,\ j = 1,2.$$

Since t will also be a continuity point of $J_j^*(\cdot, \omega)$ for $j = 1,2$, then by (8.42), $\hat{J}_j^n(t, \omega) \to J_j^*(t, \omega)$ as $n \to \infty$ for $j = 1,2$ and, hence, by (8.45),

$$(8.46) \qquad \hat{K}_j^n(t, \omega) \to J_j^*(t, \omega) \quad \text{as } n \to \infty \quad \text{for } j = 1,2.$$

Since $\hat{K}_j^n(\cdot,\omega)$ and $J_j^*(\cdot,\omega)$ are nonnegative and nondecreasing functions, and the set of t's for which (8.46) holds is dense and includes $t = 0$, it follows from Proposition 2.1(iii) that $\hat{K}_j^n(\cdot,\omega) \to \hat{J}_j^*(\cdot,\omega)$ in the $\mathbf{M}_1$ topology as $n \to \infty$, for $j = 1, 2$. Thus, as $n \to \infty$,

$$(\hat{W}^n, \hat{K}_1^n, \hat{K}_2^n)(\cdot,\omega) \to (W^*, J_1^*, J_2^*)(\cdot,\omega), \tag{8.47}$$

with the product topology on $D^3 = D \times D \times D$, where the first copy of D has the $\mathbf{J}_1$ topology and the other two copies each have the $\mathbf{M}_1$ topology. Now by Lemma 7.2(iii), the sets of times of discontinuity of $J_1^*(\cdot,\omega)$ and $J_2^*(\cdot,\omega)$ are disjoint and $W^*(\cdot,\omega)$ is continuous. Thus, it follows from the continuity of addition under these conditions [see Proposition 2.1(ii)] that as $n \to \infty$,

$$\begin{aligned}\hat{V}_1^n(\cdot,\omega) &\equiv (2\hat{W}^n - \hat{K}_1^n + \hat{K}_2^n)(\cdot,\omega)\\ &\to (2W^* - J_1^* + J_2^*)(\cdot,\omega) \equiv U_1^*(\cdot,\omega)\end{aligned} \tag{8.48}$$

in the $\mathbf{M}_1$ topology. Hence, a.s. as $n \to \infty$,

$$\hat{V}_1^n \to U_1^* \quad \text{in the } \mathbf{M}_1 \text{ topology.} \tag{8.49}$$

Combining this with (8.30) and (8.33), we have that a.s. as $n \to \infty$,

$$(\hat{V}_1^n, \hat{V}_2^n) \to (U_1^*, U_2^*) \tag{8.50}$$

with the product topology on $D^2 = D \times D$, where the first copy of D has the $\mathbf{M}_1$ topology and the second copy of D has the $\mathbf{J}_1$ topology.

Turning to the idleness processes, we have by (5.4), (6.30), (8.16), (8.17), (8.35), (8.38) and (8.40),

$$\begin{aligned}\hat{I}_1^n(t) &= \frac{1}{\sqrt{n}} I_1^n(nt) = \sqrt{n}\int_0^t 1_{\hat{R}^n}(\hat{U}^n(s))\,ds\\ &= \hat{P}_1^n(t) + \frac{1}{2}\hat{K}_1^n(t),\end{aligned} \tag{8.51}$$

where

$$\hat{P}_1^n(t) = \int_0^t 1_{\hat{R}^n}(\hat{U}^n(s-))\,d\big(\sqrt{n}\,s - \tfrac{1}{2}\hat{B}_1^n(s)\big). \tag{8.52}$$

By Lemma 6.1 and stochastic integration, $\hat{P}_1^n$ is a martingale relative to the filtration $\{\mathscr{G}_{nt}, t \ge 0\}$. Then by Doob's L^2 maximal inequality, the L^2 isometry for stochastic integrals ([17], pages 66–68) and the martingale property of $\frac{1}{2}\hat{B}_1^n(t) - \sqrt{n}\,t$, which has quadratic variation $(4\sqrt{n})^{-1}\hat{B}_1^n(t)$, we have

$$\begin{aligned}E\Big[\sup_{0\le s\le t}|\hat{P}_1^n(s)|^2\Big] &\le 4E\big[|\hat{P}_1^n(t)|^2\big]\\ &= \frac{1}{\sqrt{n}}E\Big[\int_0^t 1_{\hat{R}^n}(\hat{U}^n(s-))\,d\hat{B}_1^n(s)\Big]\\ &= \frac{1}{\sqrt{n}}E\big[\hat{K}_1^n(t)\big].\end{aligned} \tag{8.53}$$

By (8.47), a.s. $\hat{K}_1^n(t) \to J_1^*(t)$ at all continuity points t of J_1^*. It follows (cf. [1], page 124) that for all but countably many t (not depending on ω), $\hat{K}_1^n(t) \to J_1^*(t)$ a.s. and hence $(1/\sqrt{n})\hat{K}_1^n(t) \to 0$ a.s. Furthermore, for each t, $\{(1/\sqrt{n})\hat{K}_1^n(t)\}_{n=1}^\infty$ is uniformly integrable since for all n,

$$\begin{aligned}
E\left[\left(\frac{1}{\sqrt{n}}\hat{K}_1^n(t)\right)^2\right] &= 4E\left[\left(\int_0^t 1_{\hat{R}^n}\big(\hat{U}^n(s)\big)\,ds - \frac{1}{\sqrt{n}}\hat{P}_1^n(t)\right)^2\right]\\
&\le 8E\left[\left(\int_0^t 1_{\hat{R}^n}\big(\hat{U}^n(s)\big)\,ds\right)^2\right] + \frac{8}{n}E\left[\big(\hat{P}_1^n(t)\big)^2\right]\\
&\le 8t^2 + \frac{2}{n^{3/2}}E\left[\int_0^t 1_{\hat{R}^n}\big(\hat{U}^n(s-)\big)\,d\hat{B}_1^n(s)\right] \qquad (8.54)\\
&\le 8t^2 + \frac{4}{n}E\left[\int_0^t 1_{\hat{R}^n}\big(\hat{U}^n(s)\big)\,ds\right]\\
&\le 8(t^2+t),
\end{aligned}$$

where we have used (8.51) to rewrite $\hat{K}_1^n(t)$, used part of (8.53) to rewrite the mean of $(\hat{P}_1^n(t))^2$ and used the fact that $\hat{B}_1^n(t) - 2\sqrt{n}\,t$ defines a martingale to evaluate it. It follows that for all but countably many t,

$$\frac{1}{\sqrt{n}}E\big[\hat{K}_1^n(t)\big] \to 0 \quad \text{as } n \to \infty. \tag{8.55}$$

Combining (8.51)–(8.55) yields that $\{\sup_{0\le s\le t}|\hat{I}_1^n(s) - \frac{1}{2}\hat{K}_1^n(s)|\}_{n=1}^\infty$ converges to zero in L^2 for each $t \ge 0$. The same result holds with the subscript 2 in place of 1. It follows from this and the a.s. convergence expressed in (8.47) that

$$\big(\hat{I}_1^n, \hat{I}_2^n\big) \to \tfrac{1}{2}(J_1^*, J_2^*) \quad \text{in probabiity as } n \to \infty, \tag{8.56}$$

where $D^2 = D \times D$ has the product topology in which each copy of D has the $\mathbf{M}_1$ topology.

Combining (5.2)–(5.4), (3.3)–(3.6), (8.31), (8.50) and (8.56), we have, as $n \to \infty$,

$$\begin{aligned}
\hat{Q}_2^n(\cdot) &= \frac{1}{\sqrt{n}}Q_2^n(n\,\cdot) = \big[\hat{V}_1^n(\cdot)\big]^+ \to \big[U_1^*(\cdot)\big]^+ \quad \text{a.s. in the } \mathbf{M}_1 \text{ topology},\\
\hat{Q}_4^n(\cdot) &= \frac{1}{\sqrt{n}}Q_4^n(n\,\cdot) = \big[\hat{V}_1^n(\cdot)\big]^- \to \big[U_1^*(\cdot)\big]^- \quad \text{a.s. in the } \mathbf{M}_1 \text{ topology},\\
\hat{Q}_1^n(\cdot) &= \frac{1}{n}Q_1^n(n\,\cdot) = \hat{V}_2^n(\cdot) - \frac{1}{\sqrt{n}}\hat{Q}_2^n(\cdot) \to U_2^*(\cdot) \quad \text{a.s. in the } \mathbf{J}_1 \text{ topology},
\end{aligned}$$

$$\hat{Q}_3^n(\cdot) = \frac{1}{n}Q_3^n(n\,\cdot) = 1 - \left(\hat{Q}_1^n(\cdot) + \frac{1}{\sqrt{n}}\hat{Q}_2^n(\cdot) + \frac{1}{\sqrt{n}}\hat{Q}_4^n(\cdot)\right)$$
$$\to 1 - U_2^*(\cdot) \quad \text{a.s. in the } \mathbf{J}_1 \text{ topology},$$
$$\hat{I}_j^n(\cdot) = \frac{1}{\sqrt{n}}I_j^n(n\,\cdot) \to \frac{1}{2}J_j^* \quad \text{in probability for the } \mathbf{M}_1 \text{ topology}, j = 1, 2.$$

It follows that for Q^* and I^* defined by (5.8)–(5.10), $(\hat{Q}^n, \hat{I}^n) \to (Q^*, I^*)$ in probability as $n \to \infty$, where D^6 has the product topology described in Theorem 5.1. Thus Theorem 5.1 follows.

In fact, the limit processes Q_1^* and Q_3^* are continuous and so the weak convergence $(\hat{Q}_1^n, \hat{Q}_3^n) \Rightarrow (Q_1^*, Q_3^*)$ is in the $\mathbf{J}_1$ topology on D^2 [cf. (8.9)]. Furthermore, the mapping $x \to (x^+, x^-)$ from D with the $\mathbf{M}_1$ topology into D^2 with the $\mathbf{M}_1$ topology is continuous, and $(\hat{Q}_2^n, \hat{Q}_4^n), (Q_2^*, Q_4^*)$ are defined by applying this mapping to $\hat{V}_1^n, U_1^*$, respectively. Thus, it follows from (8.49) and the continuous mapping theorem that $(\hat{Q}_2^n, \hat{Q}_4^n) \Rightarrow (Q_2^*, Q_4^*)$ in the $\mathbf{M}_1$ topology on D^2.

Corollary 5.2 follows from Theorem 5.1 plus the realization that the convergence in the product topology there can be replaced by that in the $\mathbf{J}_1$ topology on D^6 because all of the limit processes are continuous.

APPENDIX A

Proofs for decomposition of the limit process.

Proof of Lemma 7.1. For part (i), let $T_1 = \inf\{t \ge 0\colon Z^*(t) \in \Delta^*\}$. If $T_1 = \infty$, then the desired result is clearly true. So we assume $T_1 < \infty$ and by symmetry we may suppose that $Z^*(T_1) \in \Phi^*$. Let $S_1 = \inf\{t \ge T_1\colon Z^*(t) = (0,0)\}$. Fix $\delta > 0$. Let $\alpha_1 = \inf\{t \ge T_1\colon Z_1^*(t) < -\delta\}$. Since Z^* sticks to Φ^* until it reaches $(0,0)$ and Z_1^* is a one-dimensional Brownian motion, we have a.s. $S_1 \le \alpha_1 < \infty$. Let $\gamma_1 = \inf\{t \ge \alpha_1\colon Z^*(t) \in \Psi^* \text{ or } Z_1^*(t) = 0\}$. Then for $\alpha_1 \le t < \gamma_1$, $dZ_2^*(t) = 2\,dt$ and $Z_1^*(t)$ behaves like a one-dimensional Brownian motion. It follows that $\gamma_1 < \infty$ a.s. and there is $\varepsilon > 0$ such that

$$\inf_{x\colon x_1 \le -\delta} P\bigl(Z^*(\gamma_1) \in \Psi^* \big| Z^*(\alpha_1) = x\bigr)$$
$$\ge P\bigl(Z_1^*\bigl(\alpha_1 + \tfrac{1}{2}\bigr) - Z_1^*(\alpha_1) < \delta\bigr) = \varepsilon > 0.$$

Similarly, if $\alpha_2 = \inf\{t \ge \gamma_1\colon Z_1^*(t) \ge \delta\}$ and $\gamma_2 = \inf\{t \ge \alpha_2\colon Z^*(t) \in \Phi^*$ or $Z_1^*(t) = 0\}$, then $\alpha_2 \le \gamma_2 < \infty$ a.s. and

$$\inf_{x\colon x_1 \ge \delta} P\bigl(Z^*(\gamma_2) \in \Phi^* \big| Z^*(\alpha_2) = x\bigr) \ge \varepsilon > 0.$$

One can now use a regeneration argument to show that $T_2 \equiv \inf\{t \ge S_1\colon Z^*(t) \in \Psi^*\} < \infty$ a.s. Since the quickest way for Z^* to go from $(0,0)$ to Ψ^* is to drift upward at rate 2, it follows that

$$\int_{S_1}^{T_2} 1_{\Lambda^*}(Z^*(s))\,ds \ge \tfrac{1}{2}.$$

Let $S_2 = \inf\{t \geq T_2\colon Z^*(t) = (0,1)\}$. Continuing in this manner, defining T_{n+1}, S_{n+1} for $n \geq 2$, associated in the obvious way with alternating visits to the boundary segments Φ^*, Ψ^*, we obtain a.s.

$$\lambda^*(\infty) \geq \sum_{n=1}^{\infty} \int_{S_n}^{T_{n+1}} 1_{\Lambda^*}(Z^*(s))\, ds \geq \sum_{n=1}^{\infty} \tfrac{1}{2} = \infty.$$

For part (ii), observe that

$$\lambda^*(t) \geq \int_0^{t \wedge \gamma} 1_{(0,\infty)}(Z_1^*(s))\, ds,$$

where $\gamma = \inf\{s \geq 0\colon Z^*(s) \in \Phi^*\}$. Since $Z^*(0) = (0,1)$ and Z^* has continuous paths, the stopping time $\gamma > 0$. Now Z_1^* is a one-dimensional Brownian motion and so $\int_0^t 1_{(0,\infty)}(Z_1^*(s))\, ds > 0$ for all $t > 0$ a.s. The desired result then follows.

We note that the statements about τ^* in Lemma 7.1 follow immediately from the properties of λ^* and the definition of τ^* as the right continuous inverse of λ^*. □

Proof of Lemma 7.2. We shall prove the properties of J_2^*; the proof is analogous for J_1^*. We first prove (i). For each $m \geq 1$, let

$$\Psi_m^* = \left\{ z^* \in S^*\colon z_1^* \leq -\frac{1}{m},\, z_2^* = 1 \right\}$$

and define a sequence of pairs of stopping times $\{(\gamma_m^n, \delta_m^n)\}_{n=0}^{\infty}$ such that

$$\begin{aligned}
\gamma_m^0 &= \inf\{t \geq 0\colon Z^*(t) \in \Psi_m^*\},\\
\delta_m^0 &= \inf\{t \geq \gamma_m^0\colon Z_1^*(t) = 0\},\\
\gamma_m^n &= \inf\{t \geq \delta_m^{n-1}\colon Z^*(t) \in \Psi_m^*\}, \qquad n \geq 1,\\
\delta_m^n &= \inf\{t \geq \gamma_m^n\colon Z_1^*(t) = 0\}, \qquad n \geq 1.
\end{aligned}$$

Let $\Gamma_m = \bigcup_{n=0}^{\infty} [\gamma_m^n, \delta_m^n)$ and

$$N_{2,m}^*(t) = \int_0^t 1_{\Gamma_m}(s)\, dX_1^*(s).$$

For $s \in \Gamma_m$, $Z^*(s) \in \Psi^*$ and for each $t \geq 0$, $\int_0^t 1_{\Gamma_m}(s)\, ds \to \int_0^t 1_{\Psi^*}(Z^*(s))\, ds$ a.s. as $m \to \infty$. It then follows from the L^2-isometry of stochastic calculus ([17], pages 66–68) that for each $t \geq 0$, $N_{2,m}^*(t) \to N_2^*(t)$ in L^2 as $m \to \infty$, and hence (cf. Theorem 2.6 of [5]) there is a subsequence $\{N_{2,m_k}^*\}_{k=1}^{\infty}$ that a.s. converges u.o.c. to N_2^*. Hence, since $\tau^*(t) < \infty$ a.s., $\{N_{2,m_k}^*(\tau^*(t))\}_{k=1}^{\infty}$ converges in probability to $J_2^*(t)$ for each $t \geq 0$. Thus, if we show that for each m, $N_{2,m}^*(\tau^*(\cdot))$ is nondecreasing a.s., it will follow that this property is inherited by J_2^*.

For $s \in [\gamma_m^n, \delta_m^n)$, $Z^*(s) \in \Psi^*$ and $\lambda^*(s) = \lambda^*(\gamma_m^n)$. It follows that $\tau^*(t) \notin \Gamma_m$ for all $t \geq 0$. Hence, a.s. for all $t \geq 0$,

$$N_{2,m}^*(\tau^*(t)) = \sum_{n=0}^{\infty} (X_1^*(\delta_m^n) - X_1^*(\gamma_m^n)) 1_{\{\delta_m^n \leq \tau^*(t)\}}$$

which is clearly an a.s. nondecreasing process, since $\tau^*(\cdot)$ is nondecreasing, $X_1^*(\delta_m^n) = Z_1^*(\delta_m^n) = 0$ and $X_1^*(\gamma_m^n) < 0$ on $\{\delta_m^n \le \tau^*(t)\} \subset \{\delta_m^n < \infty\}$ a.s. This completes the proof that J_2^* is a.s. nondecreasing.

Property (ii) follows immediately from the facts that $\tau^*(0) = 0$ a.s. [see Lemma 7.1(ii)] and $N_2^*(0) = 0$.

For property (iii), we argue sample path by sample path. For this, consider a realization of J_2^* (and corresponding realizations of U^*, Z^* and τ^*). Suppose that t is a point of increase of J_2^* and for a contradiction, suppose that $U^*(t) \neq (0,1)$. Then $U^*(t) \equiv Z^*(\tau^*(t)) \in \Lambda^* \setminus \{(0,1)\}$. It follows from the continuity of the paths of Z^*, the fact that τ^* only deletes the time that Z^* is in $\Phi^* \cup \Psi^*$ and Lemma 7.1(i) that there is an $\varepsilon > 0$ such that $Z^*(s) \notin \Psi^*$ for all $s \in [\tau^*(t-\varepsilon), \tau^*(t+\varepsilon)]$. Now

$$J_2^*(t+\varepsilon) - J_2^*(t-\varepsilon) = N_2^*(\tau^*(t+\varepsilon)) - N_2^*(\tau^*(t-\varepsilon)).$$

The latter is zero for all except a null set (not depending on t) of realizations because a.s. N_2^* does not change while Z^* is in $S^* \setminus \Psi^*$. This yields the desired contradiction. □

Proof of Lemma 7.3. For $G \subset \mathbb{R}^n$ and any integer $k \ge 1$, let $C_b^k(G)$ denote the space of real-valued functions that have derivatives up to and including order k on some domain containing G and which together with these derivatives are continuous and bounded on G.

We shall first prove that for each $t \ge 0$,

$$E\left[\int_0^t 1_{\Lambda_1^* \setminus \{(0,0)\}}(Z^*(s))\, d(s + Y_1^*(s))\right] = 0. \tag{A.1}$$

A similar argument can be used to prove that (A.1) holds with $\Lambda_2^* \setminus \{(0,1)\}, Y_2^*$ in place of $\Lambda_1^* \setminus \{(0,0)\}, Y_1^*$, respectively.

Fix $\delta > 0$ and $\varepsilon > 0$. Let $g \in C_b^2(\mathbb{R})$ and $h \in C_b^2([0,1])$ such that g is nonincreasing, $g \equiv 1$ on $(-\infty, -2\delta]$, $g \equiv 0$ on $[-\delta, \infty)$, h' is nonincreasing, $h' \equiv 1$ on $[0, \varepsilon]$, $h' \equiv 0$ on $[2\varepsilon, 1]$ and $h(x) \equiv \int_0^x h'(u)\, du$. In particular, $|h(x)| \le 2\varepsilon$. Define $f(z) = g(z_1)h(z_2)$ for all $z = (z_1, z_2) \in S^*$. Then by Itô's formula we have a.s. for each $t \ge 0$,

$$\begin{aligned} f(Z^*(t)) &- f(Z^*(0)) \\ &= \int_0^t g'(Z_1^*(s))h(Z_2^*(s))\, dZ_1^*(s) \\ &\quad + 2\int_0^t g(Z_1^*(s))h'(Z_2^*(s))\, ds \\ &\quad + \int_0^t g(Z_1^*(s))h'(0)\, dY_1^*(s) \\ &\quad + 2\int_0^t g''(Z_1^*(s))h(Z_2^*(s))\, ds. \end{aligned} \tag{A.2}$$

Here we have used the fact that $g \equiv 0$ on $[-\delta, \infty)$ and $h'(1) = 0$ to simplify the integrals involving the drift and the boundary controls for Z_2^*. In (A.2), the stochastic integral with respect to the Brownian motion Z_1^* is a martingale relative to the filtration generated by Z^* and so taking expectations in (A.2) yields

$$\begin{aligned} &E[f(Z^*(t)) - f(Z^*(0))] - 2E\left[\int_0^t g''(Z_1^*(s))h(Z_2^*(s))\,ds\right] \\ &\quad = 2E\left[\int_0^t g(Z_1^*(s))h'(Z_2^*(s))\,ds\right] + E\left[\int_0^t g(Z_1^*(s))\,dY_1^*(s)\right] \\ &\quad \geq 2E\left[\int_0^t 1_{(-\infty,\,-2\delta]}\,(Z_1^*(s))1_{[0,\,\varepsilon]}(Z_2^*(s))\,ds\right] \\ &\qquad + E\left[\int_0^t 1_{(-\infty,\,-2\delta]}\,(Z_1^*(s))\,dY_1^*(s)\right]. \end{aligned}$$

Now the left-hand member above is dominated by $4\varepsilon + 4\varepsilon t \max\{|g''(x)|: x \in [-2\delta, -\delta]\}$ and so letting $\varepsilon \downarrow 0$ and then $\delta \downarrow 0$ we obtain (A.1) by Fatou's lemma (recall that Y_1^* can only increase when Z_2^* is zero).

Note that since Z_1^* is a one-dimensional Brownian motion, the following is immediate:

$$\text{(A.3)} \qquad \int_0^\infty 1_{\{0\}}(Z_1^*(s))\,ds = 0 \quad \text{a.s.}$$

Furthermore, X_2^* is a pure drift process. For such a continuous process, the two-sided reflection mapping can be obtained by piecing together one-sided reflection mappings using a sequence of stopping times. From this and the explicit nature of the one-sided reflection mapping (cf. [5], Lemma 8.1), it follows that $dY_1^*(s)$ and $dY_2^*(s)$ are absolutely continuous with respect to ds. Combining the above statements yields

$$\text{(A.4)} \qquad \int_0^\infty 1_{\{0\}}(Z_1^*(s))\,d(Y_1^* + Y_2^*)(s) = 0 \quad \text{a.s.}$$

Putting all of the above results together yields (7.14)–(7.16). □

Proof of Lemma 7.4. For each $m \geq 1$, let

$$\Delta_m^* = \left\{z^* \in S^*: z_1^* \geq \frac{1}{m} \text{ and } z_2^* = 0, \text{ or } z_1^* \leq -\frac{1}{m} \text{ and } z_2^* = 1\right\}$$

and define a sequence of pairs of stopping times $\{(\gamma_m^n, \delta_m^n)\}_{n=0}^\infty$ as in the proof of Lemma 7.2 but with Δ_m^* in place of Ψ_m^* there. Then in a similar manner to that in Lemma 7.2, except that we have pathwise integrals here rather than stochastic integrals, we have a.s. for all $t \geq 0$,

$$\text{(A.5)} \quad \int_0^t 1_{\Delta^*}(Z^*(s))\,dZ_2^*(s) = \lim_m \sum_{n=0}^\infty (Z_2^*(\delta_m^n \wedge t) - Z_2^*(\gamma_m^n \wedge t)).$$

Now Z_2^* sticks to the boundary (either $Z_2^* = 0$ or $Z_2^* = 1$) on $[\gamma_m^n, \delta_m^n)$ and so the summands in the right-hand member above are all zero a.s. Hence (7.18) holds. Combining this with (7.17) and (7.1), we obtain (7.19). □

APPENDIX B

Key results for the heavy traffic limit theorem. In this section, the centered dot $(\cdot)$ will denote a generic time t in $[0, \infty)$. In particular, if x^n, $x \in D^m$, then $x^n(\cdot) \to x(\cdot)$ u.o.c. will mean that x^n converges uniformly on compact time intervals to x as $n \to \infty$.

PROOF OF LEMMA 8.1. By the Skorokhod representation theorem (cf. [8], Theorem 3.1.8) and since all of the limit processes in (8.9) are continuous, we may assume that the convergence there is almost surely u.o.c. Given this, we shall prove that a.s.,

$$(\mathrm{B.1}) \qquad \int_0^{\cdot} 1_{(-\infty,0)}\big(\hat{X}_1^n(s-)\big)\, d\bar{A}_1^n(s) \to 2\int_0^{\cdot} 1_{(-\infty,0)}\big(X_1^*(s)\big)\, ds \quad \text{u.o.c.}$$

One can similarly show that the same result holds with $(0, \infty)$, $\bar{A}_2^n$, in place of $(-\infty, 0)$, $\bar{A}_1^n$, respectively. Then we can subtract these limit results and combine the result with (8.4) and (7.1) to obtain a.s. $\hat{X}_2^n \to X_2^*$ u.o.c. Lemma 8.1 follows from this and the almost sure convergence assumed at the beginning of this proof.

Proposition B.1 below is used for the proof of (B.1). This proposition follows by essentially the same proof as that for Lemma 2.4 in Dai and Williams [7]. (The fact that our α^n are in D, rather than in C as they would be in [7], does not affect the validity of the proof, provided one replaces s by $s-$ in the integrands and observes that α^n converges to α u.o.c. since α is continuous.) The proposition is stated in slightly greater generality than is needed here and the reader is referred to [7] for details of the proof.

PROPOSITION B.1. *Let $m \geq 1$ and consider D^m and D to be endowed with their $\mathbf{J}_1$ topologies. Suppose that $\xi^n \to \xi$ in D^m and $\alpha^n \to \alpha$ in D. Assume α^n is nondecreasing for each n and that α is continuous. Then, for any $f \in C_b(\mathbb{R}^m)$,*

$$(\mathrm{B.2}) \qquad \int_{[0,\cdot]} f(\xi^n(s-))\, d\alpha^n(s) \to \int_{[0,\cdot]} f(\xi(s))\, d\alpha(s) \quad \textit{u.o.c.}$$

Continuing now with the proof of (B.1), we have from Proposition B.1 and the a.s. convergence assumed at the beginning of the proof of Lemma 8.1, it follows that for any fixed $f \in C_b(\mathbb{R})$, a.s.,

$$(\mathrm{B.3}) \qquad \int_0^{\cdot} f\big(\hat{X}_1^n(s-)\big)\, d\bar{A}_1^n(s) \to 2\int_0^{\cdot} f\big(X_1^*(s)\big)\, ds \quad \text{u.o.c.}$$

It remains to show that we can replace f by $1_{(-\infty,0)}$ in the above. This follows from the property of the one-dimensional Brownian motion X_1^* that

$$\int_0^\infty 1_{\{0\}}\big(X_1^*(s)\big)\,ds = 0 \quad \text{a.s.} \tag{B.4}$$

Indeed, if for each $\varepsilon > 0$, f_ε is a continuous nonincreasing function on $\mathbb{R}$ such that $f_\varepsilon = 1$ on $(-\infty, -\varepsilon]$ and $f_\varepsilon = 0$ on $[0, \infty)$, and if g_ε is a continuous function on $\mathbb{R}$ such that $0 \le g_\varepsilon \le 1$, $g_\varepsilon = 1$ on $(-\varepsilon, 0]$ and $g_\varepsilon = 0$ on $(-\infty, -2\varepsilon) \cup (\varepsilon, \infty)$, then

$$\begin{aligned}
&\left|\int_0^{\cdot} 1_{(-\infty,0)}\big(\hat{X}_1^n(s-)\big)\,d\bar{A}_1^n(s) - 2\int_0^{\cdot} 1_{(-\infty,0)}\big(X_1^*(s)\big)\,ds\right| \\
&\quad\le \left|\int_0^{\cdot} f_\varepsilon\big(\hat{X}_1^n(s-)\big)\,d\bar{A}_1^n(s) - 2\int_0^{\cdot} f_\varepsilon\big(X_1^*(s)\big)\,ds\right| \\
&\qquad + \left|\int_0^{\cdot} g_\varepsilon\big(\hat{X}_1^n(s-)\big)\,d\bar{A}_1^n(s) - 2\int_0^{\cdot} g_\varepsilon\big(X_1^*(s)\big)\,ds\right| \\
&\qquad + 4\left|\int_0^{\cdot} g_\varepsilon\big(X_1^*(s)\big)\,ds\right|.
\end{aligned} \tag{B.5}$$

Since $g_\varepsilon \to 1_{\{0\}}$ as $\varepsilon \to 0$, it follows from dominated convergence and (B.4) that a.s. the last term in (B.5) tends to zero u.o.c. as $\varepsilon \to 0$. Furthermore, for each fixed $\varepsilon > 0$, by (B.3), almost surely the remaining terms in the last member of (B.5) tend to zero u.o.c. as $n \to \infty$. The desired result (B.1) follows. □

Proof of Lemma 8.2. We first prove that

$$\begin{aligned}
&\left(\hat{Z}_1^n, \hat{Z}_2^n, \hat{X}_1^n, \hat{X}_2^n, \bar{A}_1^n, \hat{A}_1^n, \bar{A}_2^n, \hat{A}_2^n, \hat{\lambda}^n, \hat{\phi}^n, \hat{\psi}^n\right) \\
&\qquad \Rightarrow \left(Z_1^*, Z_2^*, X_1^*, X_2^*, A^*, B_1^*, A^*, B_2^*, \lambda^*, \phi^*, \psi^*\right)
\end{aligned} \tag{B.6}$$

in the $\mathbf{J}_1$ topology.

Since we have (8.11) and $\hat{\lambda}^n, \hat{\phi}^n, \hat{\psi}^n$ are all Lipschitz continuous with Lipschitz constant bounded by 1, we immediately have tightness of the sequence in the left-hand member of (B.6). Thus, it suffices to show that any weak limit point (along a subsequence) of the left member has the same distribution as the right member of (B.6).

For this, suppose a subsequence of the left member of (B.6) converges weakly to a limit process. To minimize notation, we use the same notation for this subsequence as for the original sequence and by the Skorokhod representation theorem we may assume the convergence is a.s. In view of (8.11), the first eight components of the limit have the same joint distribution as the right member of (8.11). Thus, if we take the limit process to have its first eight components given by the right member of (8.11), it suffices to show that given

$$\left(\hat{\lambda}^n, \hat{\phi}^n, \hat{\psi}^n\right) \to \left(\tilde{\lambda}, \tilde{\phi}, \tilde{\psi}\right) \quad \text{u.o.c. almost surely,}$$

we have $(\tilde{\lambda}, \tilde{\phi}, \tilde{\psi}) = (\lambda^*, \phi^*, \psi^*)$ a.s., where $\lambda^*, \phi^*, \psi^*$ are determined from Z^* by (7.8) and (8.24). Indeed, since $(\hat{\lambda}^n + \hat{\phi}^n + \hat{\psi}^n)(t) = (\lambda^* + \phi^* + \psi^*)(t) = t$, it suffices to prove the identity of any two of the components. We give the detailed proof that $\tilde{\lambda} = \lambda^*$ a.s. and sketch the similar idea for the proof that $\tilde{\phi} = \phi^*$ a.s.

By the a.s. convergence assumed above and since Z^* has continuous paths, we have that a.s. $\hat{Z}^n \to Z^*$ u.o.c. and hence for any $f \in C_b(S^*)$, a.s.,

$$f(\hat{Z}^n) \to f(Z^*) \quad \text{u.o.c.} \tag{B.7}$$

For $\varepsilon > 0$, let $f_\varepsilon \in C_b(\mathbb{R})$ such that $0 \le f_\varepsilon \le 1$, $f_\varepsilon = 1$ on $[\varepsilon, 1 - \varepsilon]$ and $f_\varepsilon = 0$ on $\mathbb{R} \setminus (0, 1)$. Then,

$$\begin{aligned} \left|\hat{\lambda}^n(\cdot) - \lambda^*(\cdot)\right| &\le \int_0^\cdot \left|1_{\Lambda^*}\big(\hat{Z}^n(s)\big) - 1_{\Lambda^*}\big(Z^*(s)\big)\right| ds \\ &\le \int_0^\cdot \left|f_\varepsilon\big(\hat{Z}_2^n(s)\big) - f_\varepsilon\big(Z_2^*(s)\big)\right| ds \\ &\quad + \int_0^\cdot \left|1_{\Lambda^*}\big(\hat{Z}^n(s)\big) - f_\varepsilon\big(\hat{Z}_2^n(s)\big)\right| ds \\ &\quad + \int_0^\cdot \left|f_\varepsilon\big(Z_2^*(s)\big) - 1_{\Lambda^*}\big(Z^*(s)\big)\right| ds. \end{aligned} \tag{B.8}$$

Note that for fixed $\varepsilon > 0$, by (B.7), a.s. the second integral in (B.8) tends to zero u.o.c. as $n \to \infty$. For the last integral, note that $f_\varepsilon \uparrow 1_{(0,1)}$ as $\varepsilon \downarrow 0$ and by (7.14), a.s. $1_{(0,1)}(Z_2^*(s)) = 1_{\Lambda^*}(Z^*(s))$ for m-a.e. s, where m denotes Lebesgue measure on $[0, \infty)$. It follows by dominated convergence that a.s.

$$\int_0^\cdot \left|f_\varepsilon\big(Z_2^*(s)\big) - 1_{\Lambda^*}\big(Z^*(s)\big)\right| ds \to 0 \quad \text{u.o.c. as } \varepsilon \to 0. \tag{B.9}$$

It remains to analyze the behavior of the third integral in (B.8). Now for $0 < \varepsilon < \delta < \frac{1}{4}$,

$$\int_0^\cdot \left|1_{\Lambda^*}\big(\hat{Z}^n(s)\big) - f_\varepsilon\big(\hat{Z}_2^n(s)\big)\right| ds \le \int_0^\cdot 1_{\Lambda_\delta \cup \Delta_{\delta,\varepsilon}}\big(\hat{Z}^n(s)\big)\, ds, \tag{B.10}$$

where

$$\begin{gathered} \Lambda_\delta = \big((-\infty, \delta] \times [0, \delta]\big) \cup \big([-\delta, \infty) \times [1 - \delta, 1]\big), \\ \Delta_{\delta,\varepsilon} = \Phi_{\delta,\varepsilon} \cup \Psi_{\delta,\varepsilon}, \\ \Phi_{\delta,\varepsilon} = [\delta, \infty) \times (0, \varepsilon], \qquad \Psi_{\delta,\varepsilon} = (-\infty, -\delta) \times [1 - \varepsilon, 1). \end{gathered}$$

Let $g_\delta \in C_b(S^*)$ such that $0 \le g_\delta \le 1$, $g_\delta = 1$ on Λ_δ and $g_\delta = 0$ on

$$\begin{aligned} &\big([2\delta, \infty) \times [0, 1 - 2\delta]\big) \cup \big((-\infty, -2\delta] \times [2\delta, 1]\big) \\ &\qquad \cup \big((-\infty, \infty) \times [2\delta, 1 - 2\delta]\big). \end{aligned}$$

Now, for δ fixed,

$$\int_0^\cdot 1_{\Lambda_\delta}\big(\hat{Z}^n(s)\big)\, ds \le \int_0^\cdot g_\delta\big(\hat{Z}^n(s)\big)\, ds, \tag{B.11}$$

where by (B.7), we have a.s.

$$\int_0^{\cdot} g_\delta(\hat{Z}^n(s))\,ds \to \int_0^{\cdot} g_\delta(Z^*(s))\,ds \quad \text{u.o.c.} \tag{B.12}$$

Furthermore, by dominated convergence, a.s.

$$\int_0^{\cdot} g_\delta(Z^*(s))\,ds \to \int_0^{\cdot} 1_{\Lambda_1^* \cup \Lambda_2^*}(Z^*(s))\,ds \quad \text{u.o.c. as } \delta \downarrow 0, \tag{B.13}$$

where the right member above is a.s. identically zero by (7.14).

Now fix $\delta \in (0, \frac{1}{4})$ and consider $n > 16/\delta^2$. We shall study $\int_0^{\cdot} 1_{\Phi_{\delta,\varepsilon}}(\hat{Z}^n(s))\,ds$. The integral with $\Psi_{\delta,\varepsilon}$ in place of $\Phi_{\delta,\varepsilon}$ can be analyzed by a symmetric argument. For $0 < \varepsilon < \delta$, let

$$h(z_1, z_2) = k(z_1) l(z_2),$$

where $k \in C_b^2(\mathbb{R})$, $l \in C_b^1([0,1])$ such that k is nondecreasing, $k(z_1) = 0$ for $z_1 \le \delta/4$, $k(z_1) = 1$ for $z_1 \ge \delta/2$ and $|k'(z_1)| \le 8/\delta$, $|k''(z_1)| \le 100/\delta^2$ for all z_1; $l'(z_2) = 1$ for $z_2 \in [0, \varepsilon]$, $l'(z_2) = 0$ for $z_2 \ge 2\varepsilon$, l' is nonincreasing and $l(z_2) = \int_0^{z_2} l'(u)\,du$, $z_2 \in [0,1]$. In particular, $0 \le l(z_2) \le 2\varepsilon$ for all $z_2 \in [0,1]$. Now by Dynkin's formula (cf. [8], Proposition 4.1.7), we have for each $t \ge 0$,

$$E\left[h(\hat{Z}^n(t)) - h(\hat{Z}^n(0))\right] = E\left[\int_0^t (\hat{G}^n h)(\hat{Z}^n(s))\,ds\right], \tag{B.14}$$

where $\hat{G}^n$ is the infinitesimal generator for $\hat{Z}^n$. Now, $(\hat{G}^n h)(z) = 0$ for $z_1 \le 0$ [since $k(z_1) = 0$ for $z_1 \le \delta/4$ and the steps of $\hat{Z}_1^n$ are of size $1/\sqrt{n} < \delta/4$] and for z: $z_1 > 0$ and $z_2 > 0$,

$$\begin{aligned}(\hat{G}^n h)(z) = 2n\Bigg[&h\left(z_1 + \frac{1}{\sqrt{n}}, z_2\right) - h(z_1, z_2)\\ &+h\left(z_1 - \frac{1}{\sqrt{n}}, z_2\right) - h(z_1, z_2)\\ &+h\left(z_1 - \frac{1}{\sqrt{n}}, z_2 - \frac{1}{n}\right) - h\left(z_1 - \frac{1}{\sqrt{n}}, z_2\right)\Bigg],\end{aligned} \tag{B.15}$$

since the transitions of $\hat{Z}^n$ from $\{z \in S^*: z_1 > 0, z_2 > 0\}$ are at rate $4n$ and are such that $\hat{Z}_1^n$ is equally likely to move to the left or right by $1/\sqrt{n}$ and if it moves to the left, then $\hat{Z}_2^n$ also moves down by $1/n$, whereas when $\hat{Z}_1^n$ moves to the right, $\hat{Z}_2^n$ stays constant (cf. Figure 5). Now by the mean value theorem, (B.15) can be rewritten as

$$(\hat{G}^n h)(z) = 2\sqrt{n}\left(h_{z_1 z_1}(z_1^*, z_2)\tilde{z}_1\right) - 2h_{z_2}\left(z_1 - \frac{1}{\sqrt{n}}, z_2^*\right), \tag{B.16}$$

where the subscripts on h denote partial differentiation with respect to those variables and $z_1^* \in [z_1 - 1/\sqrt{n}, z_1 + 1/\sqrt{n}]$, $|\tilde{z}_1| \le 2/\sqrt{n}$, $z_2^* \in [z_2 - 1/n, z_2]$. Thus, using the positivity of k, l' and bounds on k'' and l, we have for $z_1 > 0, z_2 > 0$,

$$\text{(B.17)}\qquad \begin{aligned}(\hat{G}^n h)(z) &\le 2\sqrt{n}\,|k''(z_1^*)l(z_2)\tilde{z}_1| - 2k\left(z_1 - \frac{1}{\sqrt{n}}\right)l'(z_2^*)\\ &\le 4\cdot\frac{100}{\delta^2}\cdot 2\varepsilon - 2\cdot 1_{[\delta/2+1/\sqrt{n},\infty)}(z_1)1_{(0,\varepsilon]}(z_2).\end{aligned}$$

Similarly, for $z_1 > 0$, $z_2 = 0$,

$$(\hat{G}^n h)(z) = 2n\left[h\left(z_1 + \frac{1}{\sqrt{n}}, z_2\right) + h\left(z_1 - \frac{1}{\sqrt{n}}, z_2\right) - 2h(z_1, z_2)\right] \le \frac{800\varepsilon}{\delta^2}.$$

Since $(\hat{G}^n h)(z) = 0$ for $z_1 \le 0$, the above estimate also applies in this case. Substituting these estimates in (B.14) and rearranging yields for all $t \ge 0$,

$$\text{(B.18)}\qquad \begin{aligned}2E&\left[\int_0^t 1_{[\delta/2+1/\sqrt{n},\infty)}\left(\hat{Z}_1^n(s)\right)1_{(0,\varepsilon]}\left(\hat{Z}_2^n(s)\right)ds\right]\\ &\le E\left[h\left(\hat{Z}^n(0)\right) - h\left(\hat{Z}^n(t)\right)\right] + 800\varepsilon\delta^{-2}t\\ &\le 4\varepsilon + 800\varepsilon\delta^{-2}t.\end{aligned}$$

Thus since $1/\sqrt{n} < \delta/4 < \delta/2$,

$$\text{(B.19)}\qquad E\left[\int_0^t 1_{\Phi_{\delta,\varepsilon}}\left(\hat{Z}^n(s)\right)ds\right] \le \tfrac{1}{2}(4\varepsilon + 800\varepsilon\delta^{-2}t) \quad \text{for all } t \ge 0.$$

By symmetry, the same estimate holds with $\Psi_{\delta,\varepsilon}$ in place of $\Phi_{\delta,\varepsilon}$.

Combining the above results, we have for all $t \ge 0$,

$$\text{(B.20)}\qquad \begin{aligned}E\left[\left|\tilde{\lambda}(t) - \lambda^*(t)\right|\right] &= \lim_{n\to\infty} E\left[\left|\hat{\lambda}^n(t) - \lambda^*(t)\right|\right]\\ &\le \limsup_{n\to\infty} E\left[\int_0^t \left|1_{\Lambda^*}\left(\hat{Z}^n(s)\right) - f_\varepsilon\left(\hat{Z}_2^n(s)\right)\right|ds\right]\\ &\quad + E\left[\int_0^t \left|f_\varepsilon\left(Z_2^*(s)\right) - 1_{\Lambda^*}\left(Z^*(s)\right)\right|ds\right]\\ &\le \limsup_{n\to\infty}\left(E\left[\int_0^t g_\delta\left(\hat{Z}^n(s)\right)ds\right] + E\left[\int_0^t 1_{\Delta_{\delta,\varepsilon}}\left(\hat{Z}^n(s)\right)ds\right]\right)\\ &\quad + E\left[\int_0^t \left|f_\varepsilon\left(Z_2^*(s)\right) - 1_{\Lambda^*}\left(Z^*(s)\right)\right|ds\right]\\ &\le E\left[\int_0^t g_\delta\left(Z^*(s)\right)ds\right] + 4\varepsilon + 800\varepsilon\delta^{-2}t\\ &\quad + E\left[\int_0^t \left|f_\varepsilon\left(Z_2^*(s)\right) - 1_{\Lambda^*}\left(Z^*(s)\right)\right|ds\right],\end{aligned}$$

where we have used dominated convergence for the first line, (B.8), (B.7) and dominated convergence for the second line, (B.10) and (B.11) for the third line and (B.12) and (B.19) for the fourth line. Now by first letting $\varepsilon \downarrow 0$ and then $\delta \downarrow 0$, we conclude using (B.9), (B.13) and dominated convergence that the left member of (B.20) is zero and hence $\tilde{\lambda}(t) = \lambda^*(t)$ a.s. Since t was arbitrary and by the regularity of $\tilde{\lambda}$, λ^*, it follows that a.s. $\tilde{\lambda} = \lambda^*$.

The above proof can be modified to show that $\tilde{\phi} = \phi^*$ a.s. Essentially one chooses f_ε to equal 1 on $[\varepsilon, 1]$ and to be 0 on $[0, \varepsilon/2]$. A similar argument to that given above then yields a.s.

$$\begin{aligned} t - \hat{\phi}^n(t) &= \int_0^t 1_{S^*\setminus\Phi^*}\big(\hat{Z}^n(s)\big)\,ds \\ &\to \int_0^t 1_{S^*\setminus\Phi^*}(Z^*(s))\,ds = t - \phi^*(t) \quad \text{as } n \to \infty, \end{aligned} \tag{B.21}$$

where the convergence is uniform for t in each compact time interval. Hence $\tilde{\phi} = \phi^*$ a.s. This completes the verification of (B.6).

We now turn to verification of the full statement (8.25). By the Skorokhod representation theorem and the continuity of the limit processes, we may assume the convergence in (B.6) is almost surely u.o.c. We shall prove that for each $t \geq 0$,

$$\sup_{0 \leq s \leq t} \big|\hat{M}^n(s) - M^*(s)\big| \to 0 \quad \text{in probability as } n \to \infty. \tag{B.22}$$

The same result with $\hat{N}_j^n, N_j^*$ in place of $\hat{M}^n, M^*$, respectively, for $j = 1, 2$, can be proved in a similar manner. It follows from this and the a.s. convergence assumed for (B.6) that (8.25) holds.

For the proof of (B.22), let f_ε be as in the above proof of (B.6). Then for each $t \geq 0$,

$$\begin{aligned} \hat{M}^n(t) - M^*(t) = {} & \tfrac{1}{2}\int_0^t \Big(1_{\Lambda^*}\big(\hat{Z}^n(s-)\big) - f_\varepsilon\big(\hat{Z}_2^n(s-)\big)\Big)\,d\hat{Z}_1^n(s) \\ & + \tfrac{1}{2}\int_0^t f_\varepsilon\big(\hat{Z}_2^n(s-)\big)\,d\hat{Z}_1^n(s) - \tfrac{1}{2}\int_0^t f_\varepsilon(Z_2^*(s))\,dZ_1^*(s) \\ & + \tfrac{1}{2}\int_0^t \big(f_\varepsilon(Z_2^*(s)) - 1_{\Lambda^*}(Z^*(s))\big)\,dZ_1^*(s). \end{aligned} \tag{B.23}$$

For fixed $\varepsilon > 0$, by the a.s. uniform convergence on compacts assumed for (B.6) and the continuity of f_ε, we have a.s.

$$\big(\hat{Z}_1^n, f_\varepsilon\big(\hat{Z}_2^n\big)\big) \to (Z_1^*, f_\varepsilon(Z_2^*)) \quad \text{u.o.c.} \tag{B.24}$$

Now, $\hat{Z}_1^n$ is a pure jump martingale which moves by jumps of size $1/\sqrt{n}$ and its quadratic variation process is given (cf. [17], page 63) by

$$\begin{aligned} \big[\hat{Z}_1^n\big](t) &= \sum_{0 \leq s \leq t} \Big(\big(\Delta\hat{Z}_1^n(s)\big)^2\Big) = \sum_{0 \leq s \leq t} \Big(\big(\Delta\hat{A}_1^n(s)\big)^2 + \big(\Delta\hat{A}_2^n(s)\big)^2\Big) \\ &= \sum_{0 \leq s \leq t} \left(\frac{1}{n}\Delta A_1(ns) + \frac{1}{n}\Delta A_2(ns)\right) \\ &= \frac{1}{n}\big(A_1(nt) + A_2(nt)\big) = \bar{A}_1(t) + \bar{A}_2(t), \end{aligned} \tag{B.25}$$

where $\Delta\hat{Z}_1^n(s)$ denotes the jump of $\hat{Z}_1^n$ at s and so forth. In (B.25) we have used the fact that the A_j, $j = 1, 2$, are independent and have upward jumps of unit length. Since $\bar{A}_j(t) - 2t$ defines a martingale for $j = 1, 2$, we have

$$E\big(\big[\hat{Z}_1^n\big](t)\big) = 4t, \qquad t \geq 0. \tag{B.26}$$

It then follows from (B.24), (B.26) and Theorem 2.2 of Kurtz and Protter [14] that

$$\sup_{0 \leq u \leq t}\left|\int_0^u f_\varepsilon\big(\hat{Z}_2^n(s-)\big)\,d\hat{Z}_1^n(s) - \int_0^u f_\varepsilon(Z_2^*(s))\,dZ_1^*(s)\right| \\ \to 0 \quad \text{in probability as } n \to \infty.$$

This takes care of the second and third integrals in (B.23). For the first integral in (B.23), note that by Doob's maximal inequality for L^2 martingales, the L^2 isometry for stochastic integrals (cf. [17], pages 66–68) and the martingale property of $[\hat{Z}_1^n](t) - 4t$, we have

$$\begin{aligned} &E\left[\sup_{0 \leq u \leq t}\left|\int_0^u \big(1_{\Lambda^*}(\hat{Z}^n(s-)) - f_\varepsilon\big(\hat{Z}_2^n(s-)\big)\big)\,d\hat{Z}_1^n(s)\right|^2\right] \\ &\quad \leq 4E\left[\left|\int_0^t \big(1_{\Lambda^*}(\hat{Z}^n(s-)) - f_\varepsilon\big(\hat{Z}_2^n(s-)\big)\big)\,d\hat{Z}_1^n(s)\right|^2\right] \\ &\quad = 4E\left[\int_0^t \big|1_{\Lambda^*}(\hat{Z}^n(s-)) - f_\varepsilon\big(\hat{Z}_2^n(s-)\big)\big|^2\,d\big[\hat{Z}_1^n\big](s)\right] \\ &\quad = 16E\left[\int_0^t \big|1_{\Lambda^*}(\hat{Z}^n(s)) - f_\varepsilon\big(\hat{Z}_2^n(s)\big)\big|^2\,ds\right]. \end{aligned} \tag{B.27}$$

Now for all $z \in S^*$, in the notation of (B.10),

$$0 \leq 1_{\Lambda^*}(z) - f_\varepsilon(z_2) \leq 1_{\Lambda_\delta \cup \Delta_{\delta,\varepsilon}}(z)$$

and so combining this with (B.11) and (B.19), we see that the last member of (B.27) is dominated for $0 < \varepsilon < \delta < \frac{1}{4}$ and $n > 16/\delta^2$ by

$$16\left(E\left[\int_0^t g_\delta\big(\hat{Z}^n(s)\big)\,ds\right] + (4\varepsilon + 800\varepsilon\delta^{-2}t)\right). \tag{B.28}$$

It then follows from (B.12) and (B.13) that the $\limsup_{n\to\infty}$ of the first member of (B.27) can be made arbitrarily small, provided ε and δ are sufficiently small. Finally, for the last integral in (B.23), we note that similar manipulations to those for (B.27) yield

$$\begin{aligned} &E\left[\sup_{0 \leq u \leq t}\left|\int_0^u (f_\varepsilon(Z_2^*(s)) - 1_{\Lambda^*}(Z^*(s)))\,dZ_1^*(s)\right|^2\right] \\ &\qquad \leq 16E\left[\int_0^t |f_\varepsilon(Z_2^*(s)) - 1_{\Lambda^*}(Z^*(s))|^2\,ds\right], \end{aligned}$$

where the last member above tends to zero as $\varepsilon \to 0$, by dominated convergence [cf. (B.9)].

Combining all of the above results, we see that (B.22) holds. This completes the proof of (8.25). □

Acknowledgments. The ideas developed in this paper originated during a four-week period in 1994 when both authors were in residence at the Institute for Mathematics and its Applications (IMA) in Minneapolis. Many participants in the IMA workshop on stochastic networks contributed to its genesis, but Jim Dai, Viên Nguyen, P. R. Kumar and Gideon Weiss are particularly notable in that regard. Thanks are due Avner Friedman for management that contributed to the Institute's stimulating and collegial atmosphere, and to Maury Bramson and Bruce Hajek for insightful comments on a preliminary presentation of the results reported here. We also thank Tom Kurtz, Avi Mandelbaum, Viên Nguyen and Ward Whitt for helpful references on weak convergence and the $\mathbf{M}_1$ topology. Finally, we gratefully acknowledge the comments of an anonymous referee, especially those drawing our attention to the interesting problem of alternative scalings (see the second paragraph following Corollary 5.2).

REFERENCES

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

[2] BRÉMAUD, P. (1981). *Point Processes and Queues*. Springer, New York.

[3] CHEN, H. and MANDELBAUM, A. (1991). Leontief systems, RBV's and RBM's. In *Applied Stochastic Analysis* (M. H. A. Davis and R. J. Elliott, eds.) 1–43. Gordon and Breach, New York.

[4] CHEN, H. and MANDELBAUM, A. (1991). Stochastic discrete flow networks: diffusion approximations and bottlenecks. *Ann. Probab.* **19** 1463–1519.

[5] CHUNG, K. L. and WILLIAMS, R. J. (1990). *Introduction to Stochastic Integration*, 2nd ed. Birkhäuser, Boston.

[6] DAI, J. G. and WEISS, G. (1996). Stability and instability of fluid models for certain re-entrant lines. *Math. Oper. Res.* To appear.

[7] DAI, J. G. and WILLIAMS, R. J. (1995). Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons. *Teor. Veroyatnost. i Primenen.* **40** 3–53 (in Russian); *Theor. Probab. Appl.* **21**. To appear.

[8] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes*. Wiley, New York.

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

[10] HARRISON, J. M. and NGUYEN, V. (1993). Brownian models of multiclass queueing networks: current status and open problems. *Queueing Systems Theory Appl.* **13** 5–40.

[11] HARRISON, J. M. and NGUYEN, V. (1995). Some badly behaved closed queueing networks. In *Stochastic Networks* (F. P. Kelly and R. J. Williams, eds.) Springer, New York.

[12] HARRISON, J. M., WILLIAMS, R. J. and CHEN, H. (1990). Brownian models of closed queueing networks with homogeneous customer populations. *Stochastics Stochastics Reports* **29** 37–74.

[13] KURTZ, T. G. (1991). Random time changes and convergence in distribution under the Meyer–Zheng conditions. *Ann. Probab.* **19** 1010–1034.

[14] KURTZ, T. G. and PROTTER, P. (1991). Weak limit theorems for stochastic integrals and stochastic differential equations. *Ann. Probab.* **19** 1035–1070.

[15] LU, S. H. and KUMAR, P. R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Automat. Control* **36** 1406–1416.

[16] POMAREDE, J. L. (1976). A unified approach via graphs to Skorohod's topologies on the function space *D*. Ph.D. dissertation, Dept. Statistics, Yale Univ.

[17] PROTTER, P. (1990). *Stochastic Integration and Differential Equations*. Springer, New York.

[18] RYBKO, A. N. and STOLYAR, A. L. (1991). Ergodicity of stochastic processes describing the operation of an open queueing networks. *Problemy Peredachi Informatsii* **28** 2–26.

[19] SHARPE, M. (1988). *General Theory of Markov Processes*. Academic Press, San Diego.

[20] SKOROKHOD, A. V. (1956). Limit theorems for stochastic processes. *Theory Probab. Appl.* **1** 261–290.

[21] WHITT, W. (1980). Some useful functions for functional limit theorems. *Math. Oper. Res.* **3** 67–85.

[22] WILLIAMS, R. J. (1995). Semimartingale reflecting Brownian motions in the orthant. In *Stochastic Networks* (F. P. Kelly and R. J. Williams, eds.) Springer, New York.

GRADUATE SCHOOL OF BUSINESS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-mail: fharrison@gsb-lira.stanford.edu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
9500 GILMAN DRIVE
LA JOLLA, CALIFORNIA 92093-0112
E-mail: williams@math.ucsd.edu