Diffusion approximations for open queueing networks with service interruptions

Hong Chen
The University of British Columbia, Vancouver, B.C., Canada V6T 148

Ward Whitt
AT&T Bell Laboratories, Murray Hill, NJ 07974-0636, USA

Received 8 August 1991

This paper establishes functional central limit theorems describing the heavy-traffic behavior of open single-class queueing networks with service interruptions. In particular, each station has a single server which is alternatively up and down. There are two treatments of the up and down times. The first treatment corresponds to fixed up and down times and leads to a reflected Brownian motion, just as when there are no service interruptions, but with different parameters. To represent long rare interruptions, the second treatment has growing up and down times with the up and down times being of order n and $n^{1/2}$, respectively, when the traffic intensities are of order $1 - n^{-1/2}$. In this case we establish convergence in the Skorohod M_1 topology to a multidimensional reflection of multidimensional Brownian motion plus a multidimensional jump process.

Keywords: Service interruptions; vacations; limit theorems; heavy traffic; queueing networks; oblique reflection mapping; Skorohod topologies; jump-diffusion processes.

1. Introduction

In this paper we establish heavy-traffic limit theorems for open single-class queueing networks with service interruptions. In addition to an unlimited waiting space and the first-come first-served service discipline, each station has a single server which is alternatively up and down. When a station is down, service stops but arrivals continue; when a station comes up, service resumes where it left off. We allow the availability of these servers to depend on the basic arrival, service and routing variables; see section 3 for more details. In particular, we only require that a joint functional central limit theorem (FCLT) hold for all the basic processes. However, the easiest way to obtain such a joint FCLT is to have all the component processes be independent, invoking theorem 3.2 of Billingsley [2]; then the availability of the servers (i.e., the environment for the model) must evolve indepen-

© J.C. Baltzer AG, Science Publishers

dently of the basic arrival, service and routing variables, so that the principal case considered is of exogenous (independent) service interruptions.

By considering service interruptions, this paper extends previous work on heavy-traffic limit theorems for open queueing networks in Reiman [25], Johnson [21] and Chen and Mandelbaum [6]; see also Coffman and Reiman [8], Harrison and Williams [19] and Harrison [16].

We actually consider two different treatments of the service interruptions. The standard treatment, presented in section 5, is based on fixed up and down times, which leads to a long-run proportion of up time ν_j at each station j with $0<\nu_j<1$ and a FCLT for the cumulative up time at each station after translation. In this case we obtain a limiting reflected Brownian motion (RBM) just as without disruptions (theorem 5.2). The service interruptions lead to different parameters for this RBM. Our heavy-traffic limit in this case extends results for a single station in Fischer [12], Burman [3] and Asmussen [1]. (Burman [3] also establishes a light-traffic limit and proposes an interpolation approximation.)

The second treatment allows the up and down times to become longer as the system enters heavy traffic. In particular, we make the traffic intensities in the nth system of order $1-n^{-1/2}$. Then we let the up times be of order n and the down times be of order $\sqrt{n}$. Asymptotically, the long-run proportion of time each station is up is 1, but nevertheless the down times have a significant impact. In particular, the limit process is a multidimensional reflection of a multidimensional Brownian motion plus a jump process (theorem 4.1). The two different limits dramatically show the difference between long rare interruptions and more frequent shorter interruptions.

Our results for long up and down times in section 4 constitute network generalizations of corresponding results for a single station in section 3 of Kella and Whitt [22]; see [22] for further discussion. As in [22], a significant feature of the analysis here is the use of the Skorohod [26] M_1 topology on the function space D with time domain $(0,\infty)$ instead of the standard J_1 topology from [26] (which is used in Billingsley [2]) on the function space D with time domain $[0,\infty)$. In section 2, after defining the reflection map and the M_1 topology, we show that the fundamental oblique reflection map defined in Harrison and Reiman [17] and Reiman [25] is Lipschitz in the J_1 and M_1 topologies on D.

2. Preliminaries

In this section we discuss the oblique reflection map and the Skorohod [26] M_1 topology on D. Here we consider the space $D([0,T],\mathbb{R}^n)$. We will later state results for $D((0,\infty),\mathbb{R}^n)$. A sufficient condition for $x_n\to x$ in $D((0,\infty),\mathbb{R}^n)$ is to have $x_n\to x$ in $D([a,b],\mathbb{R}^n)$ for the restrictions to $[a,b]$ for all a and b with $0<a<b<\infty$; see section 2 of Whitt [30]. Elements of $\mathbb{R}^n$ are understood to be column vectors.

The reflection map

We use the reflection map introduced by Harrison and Reiman [17] for continuous functions and extended to D by Reiman [25]; see also Chen and Mandelbaum [4–6] and Mandelbaum [23]. Let the transpose Q^t of Q be a substochastic matrix (nonnegative with row sums less than or equal to 1) such that $Q^k \to 0$ as $k \to \infty$. (For the main results, it actually suffices for Q to be nonnegative with spectral radius less than 1. However, propositions 2.2. and 2.3 below do exploit the substochastic structure. With Markovian routing, Q^t corresponds to the routing matrix.)

The *reflection map*

$$(\psi, \phi) : D([0, T], \mathbb{R}^n) \to D([0, T], \mathbb{R}^{2n})$$

associated with Q maps x into a unique $(y, z) = (\psi(x), \phi(x))$ such that

$$z = x + (I - Q)y \geqslant 0 . \tag{2.1}$$

$$y_j \text{ is nondecreasing with } y_j(0) = 0, \quad 1 \leqslant j \leqslant J , \tag{2.2}$$

and

$$\int_0^\infty z_j(t)\, dy_j(t) = 0, \quad 1 \leqslant j \leqslant J . \tag{2.3}$$

Condition (2.3) means that y_j increases only at times $t \geqslant 0$ when $z_j(t) = 0, 1 \leqslant j \leqslant n$.

As noted by Harrison and Reiman [17], (2.1)–(2.3) is equivalent to (2.1)–(2.2) plus

$$y = \pi_x(y) \equiv (Qy - x)^\uparrow \vee 0 , \tag{2.4}$$

where $(x \vee 0) = (x_1 \vee 0, \ldots, x_n \vee 0)^t, x_1 \vee 0 = \max\{x_1, 0\}, x^\uparrow = (x_1^\uparrow, \ldots, x_n^\uparrow)^t$ and $x_1^\uparrow(t) = \sup_{0 \leqslant s \leqslant t} x_1(s), t \geqslant 0$. (Their arguments remains valid for $x \in D$.)

Harrison and Reiman [17] proved that the reflection map is well defined and continuous on C. Johnson [21], p. 67, observed that their argument extends to D with the uniform topology. As noted by Chen and Mandelbaum [4], a minor extension of the Harrison–Reiman arguments shows that the reflection map is actually Lipschitz on $D([0, T], \mathbb{R}^n)$ with the uniform topology. This in turn implies that the reflection map is Lipschitz in the Skorohod [26] J_1 and M_1 topologies, as we shall show below. This extends the elementary one-dimensional Lipschitz result on p. 62 of Whitt [27] and section 6 of Whitt [30]. For further discussion of the Lipschitz property of reflection maps, see Dupuis and Ishii [9].

To be complete and to provide explicit Lipschitz bounds, we provide additional details. For $c \in \mathbb{R}^n$, let $|c| \equiv |(c_1, \ldots, c_n)^t| = (|c_1|, \ldots, |c_n|)^t$ and let $\|c\| = \sum_{j=1}^n |c_j|$. For any $n \times n$ matrix P, let

$$\|P\| = \max_j \sum_{i=1}^n |P_{ij}| . \tag{2.5}$$

(Note that $\|P_1P_2\| \leq \|P_1\| \cdot \|P_2\|$ and $\|Pc\| \leq \|P\| \cdot \|c\|$ with these definitions.) For any $x \in D$, let $|x|$ be $\{|x(t)| : 0 \leq t \leq T\}$, i.e., $|x| = (|x_1|, \ldots, |x_n|)^t$, and let

$$\|x\| = \sup_{0 \leq t \leq T} \|x(t)\| = \sup_{0 \leq t \leq T} \sum_{j=1}^{n} |x_j(t)| . \tag{2.6}$$

Let $Q^* = \Lambda^{-1} Q \Lambda$ where Λ is diagonal and $\|Q^*\| = \alpha < 1$. (Existence is noted by Harrison and Reiman [17].) Here is the minor extension; it is a consequence of the proof in Harrison and Reiman [17].

PROPOSITION 2.1

For any x_1, x_2 in D,

$$\|\psi(x_1) - \psi(x_2)\| \leq \frac{\|\Lambda\| \cdot \|\Lambda^{-1}\|}{1 - \alpha} \|x_1 - x_2\| \tag{2.7}$$

and

$$\|\phi(x_1) - \phi(x_1)\| \leq \left(1 + \frac{\|I - Q\| \cdot \|\Lambda\| \cdot \|\Lambda^{-1}\|}{1 - \alpha}\right) \|x_1 - x_2\| . \tag{2.8}$$

Proof

From Harrison and Reiman [17],

$$\psi_{Q^*}(\Lambda x) = \Lambda \psi_Q(x) ,$$

so that

$$\begin{aligned} \|\psi_Q(x_1) - \psi_Q(x_2)\| &= \|\Lambda^{-1}\Lambda\psi_Q(x_1) - \Lambda^{-1}\Lambda\psi_Q(x_2)\| \\ &= \|\Lambda^{-1}\psi_{Q^*}(\Lambda x_1) - \Lambda^{-1}\psi_{Q^*}(\Lambda x_2)\| \\ &\leq \|\Lambda^{-1}\| \cdot \|\psi_{Q^*}(\Lambda x_1) - \psi_{Q^*}(\Lambda x_2)\| \\ &\leq \|\Lambda^{-1}\| \cdot \frac{\|\Lambda x_1 - \Lambda x_2\|}{1 - \alpha} \\ &\leq \frac{\|x_1 - x_2\|}{1 - \alpha} \cdot \|\Lambda\| \cdot \|\Lambda^{-1}\| . \end{aligned}$$

For (2.8), use $\|x_1 + x_2\| \leq \|x_1\| + \|x_2\|$. □

Note that $\|\Lambda\|$ is the maximal entry of the diagonal matrix Λ. Hence, $\|\Lambda\| \cdot \|\Lambda^{-1}\|$ can be quite large in the bound of proposition 2.1. A different bound can be obtained by exploiting the fact that $\|Q^n\| = \gamma < 1$ because the transpose of Q is stochastic. Let π_x^k be the k-fold iteration of the operator π_x defined in (2.4). We first show that π_x is an n-stage contraction operator with respect to Q. This proves that there is a unique $\psi(x)$ associated with each x. We use proposition 2.2 to estab-

lish the convergence $\pi_x^k(0) \to \psi(x)$ as $k \to \infty$ for each x and thus the bounds below in proposition 2.3.

PROPOSITION 2.2

Let $\gamma = \|Q^n\|$. For any $y_1, y_2 \in D$,

$$\|\pi_x^k(y_1) - \pi_x^k(y_2)\| \leq \|Q^k|y_1 - y_2|\| \leq \|y_1 - y_2\| \text{ for all } k \geq 1$$

and

$$\|\pi_x^k(y_1) - \pi_x^k(y_2)\| \leq \gamma\|y_1 - y_2\| \text{ for } k \geq n,$$

so that $\pi_x^k(y_1) \to \psi(x)$ as $k \to \infty$.

Proof

We proceed by induction to establish the first inequality (the second being elementary). First,

$$\begin{aligned}\|\pi_x(y_1) - \pi_x(y_2)\| &= \|(Qy_1 - x)^\uparrow \vee 0 - (Qy_2 - x)^\uparrow \vee 0\| \\ &\leq \|(Qy_1 - x)^\uparrow - (Qy_2 - x)^\uparrow\| \\ &\leq \|(Qy_1 - x) - (Qy_2 - x)\| = \|Qy_1 - Qy_2\| \\ &\leq \|Q|y_1 - y_2|\| .\end{aligned}$$

Now suppose that the relation has been established up to k. Then

$$\begin{aligned}\|\pi_x^{k+1}(y_1) - \pi_x^{k+1}(y_2)\| &= \|(Q\pi_x^k(y_1) - x)^\uparrow \vee 0 - (Q\pi_x^k(y_2) - x)^\uparrow \vee 0\| \\ &\leq \|Q\pi_x^k(y_1) - Q\pi_x^k(y_2)\| \\ &\leq \|Q|\pi_x^k(y_1) - \pi_x^k(y_2)|\| \\ &\leq \|Q^{k+1}|y_1 - y_2|\| \text{ by induction .}\end{aligned}$$

□

PROPOSITION 2.3

For any $x_1, x_2 \in D$,

$$|\psi(x_1) - \psi(x_2)| \leq (I - Q)^{-1}|x_1 - x_2|, \tag{2.9}$$

so that

$$\begin{aligned}\|\psi(x_1) - \psi(x_2)\| &\leq \|(I - Q)^{-1}\| \cdot \|x_1 - x_2\| \\ &\leq \sum_{k=0}^{\infty} \|Q^k\| \cdot \|x_1 - x_2\| \\ &\leq \frac{n}{1 - \gamma}\|x_1 - x_2\| \end{aligned} \tag{2.10}$$

and

$$\|\phi(x_1) - \phi(x_2)\| \leqslant (1 + \|I - Q\| \cdot \|(I - Q)^{-1}\|)\|x_1 - x_2\|$$
$$\leqslant \left(1 + \frac{2n}{1-\gamma}\right)\|x_1 - x_2\| . \tag{2.11}$$

Proof

As on p. 305 of Harrison and Reiman [17],

$$|\pi^{n+1}_{x_1}(0) - \pi^{n+1}_{x_2}(0)\| \leqslant |Q\pi^n_{x_1}(0) - Q\pi^n_{x_2}(0)\| + |x_1 - x_2|$$
$$\leqslant (I + Q + \ldots + Q^n)|x_1 - x_2|$$

by induction. Since $\pi^n_{x_1}(0) \to \psi(x_1)$ as $n \to \infty$ by proposition 2.2., we have (2.9). Since

$$\left\|\sum_{k=0}^{\infty} Q^k\right\| \leqslant \sum_{k=0}^{\infty} \|Q^k\| \leqslant \sum_{k=0}^{n-1} \|Q^k\| + \gamma \sum_{k=0}^{\infty} \|Q^k\| ,$$

we have

$$\sum_{k=0}^{\infty} \|Q^k\| \leqslant \frac{\sum_{k=0}^{n-1} \|Q^k\|}{1-\gamma} \leqslant \frac{n}{1-\gamma} .$$

By (2.1) and (2.9),

$$|\phi(x_1) - \phi(x_2)| = |x_1 + (I - Q)\psi(x_1) - x_2 - (I - Q)\psi(x_2)|$$
$$\leqslant |x_1 - x_2| + |(I - Q)(\psi(x_1) - \psi(x_2))| ,$$

so that

$$\|\phi(x_1) - \phi(x_2\| \leqslant \|x_1 - x_2\| + |I - Q\| \cdot \|\psi(x_1) - \psi(x_2)\|$$
$$\leqslant (1 + \|I - Q\| \cdot \|(I - Q)^{-1}\|)\|x_1 - x_2\| . \qquad \square$$

Remarks

(2.1) The bounds in propositions 2.1 and 2.3 are all tight for the special case in which $Q = 0$, i.e., Q is the matrix of all 0's. This case corresponds to a network of n queues in which each queue has only external arrivals, i.e., the departures from each queue immediately leave the network. In proposition 2.1, we can have $\Lambda = I$ and $\alpha = 0$, so that

$$\|\psi(x_1) - \psi(x_2)\| \leqslant \|x_1 - x_2\| \quad \text{and} \quad \|\phi(x_1) - \phi(x_2)\| \leqslant 2\|x_1 - x_2\| .$$

In proposition 2.3 we obtain these same bounds using the first inequality in (2.11), because $Q = 0, \gamma = 0$ and $n = 1$. To see that these bounds are tight, let $T = 1, x_1(t) = 0, 0 \leqslant t \leqslant 1$, and $x_2(t) = -I_{[1/3,1/2)}(t) + I_{[1/2,1]}(t)$, where $I_A(t)$ is the

indicator function of the set A. Then $\|x_1 - x_2\| = \|x_2\| = 1$, while $\|\phi(x_1) - \phi(x_2)\| = \|\phi(x_2)\| = 2$ and $\|\psi(x_1) - \psi(x_2)\| = \|\psi(x_2)\| = 1$.

(2.2) To see that proposition 2.3 can provide a significant improvement over proposition 2.1, consider the case of $n = 2$ and

$$Q = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

which corresponds to a network of two queues in series. For proposition 2.1, we may choose $\Lambda_{11} = x$ and $\Lambda_{22} = y$ for any positive x and y such that $\alpha = x/y < 1$. From (2.7), we obtain

$$\|\psi(x_1) - \psi(x_2)\| \leqslant \frac{1}{z(1-z)} \|x_1 - x_2\|,$$

where $z = x/y$. Hence, we may obtain any modulus greater than or equal to 4, with the modulus 4 corresponding to $z = 1/2$. On the other hand, with proposition 2.3 we have $n = 2$ and $\gamma = 0$, so that from (2.10) we obtain

$$\|\psi(x_1) - \psi(x_2)\| \leqslant 2\|x_1 - x_2\|.$$

The Skorohod topologies

To treat limit processes with discontinuous sample paths, we will work with the Skorohod [26] topologies. Moreover, here we will need the M_1 topology as well as the familiar J_1 topology discussed in Billingsley [2]. It is significant that both the Lipschitz property and continuity extend to these topologies when they are established for the uniform topology. Let (D, S) denote $D([0, T], \mathbb{R}^n)$ with topology S.

The M_1 topology is defined in terms of parametric representations of the completed graphs of the functions. For $x \in D([0, T], \mathbb{R}^n)$, the *completed graph* of x is

$$\Gamma_x = \{(s, t) : s = \alpha x(t-) + (1 - \alpha)x(t) \text{ for some } \alpha, 0 \leqslant \alpha \leqslant 1\}. \quad (2.12)$$

A parametric representation of Γ_x is a continuous function $(\tilde{x}, \tilde{t})$ mapping [0, 1] onto Γ_x such that $\tilde{t}(\cdot)$ is nondecreasing. Let $\Pi(x)$ be the set of all parametric representations of Γ_x. A metric inducing the M_1 topology is

$$d_{M_1}(x_1, x_2) = \inf_{\substack{(\tilde{x}_i, \tilde{t}_i) \in \Pi(x_i) \\ i=1,2}} \{\|\tilde{x}_1 - \tilde{x}_2\| \vee \|\tilde{t}_1 - \tilde{t}_2\|\}; \quad (2.13)$$

see Skorohod [26] and Pomarede [24]. The M_1 topology has most of the properties of the J_1 topology. For example, when x is continuous, $x_n \to x(M_1)$ if and only if $\|x_n - x\| \to 0$. Also (D, M_1) is Polish; see section II.3 of Pomarede [24].

We now observe that continuity and Lipschitz continuity extend easily from (D, U) to (D, S) for $S = J_1$ and M_1.

PROPOSITION 2.4

If $f : (D, U) \to (D, U)$ is continuous (Lipschitz with modulus K), then $f : (D, J_1) \to (D, J_1)$ and $f : (D, M_1) \to (D, M_1)$ are continuous (Lipschitz with modulus $K \vee 1$)).

Proof

The standard J_1 and M_1 metrics are characterized via the uniform metric. For J_1, from p. 111 of Billingsley [2], we obtain the Lipschitz property via

$$\begin{aligned} d_{J_1}(f(x_1), f(x_2)) &= \inf_{\lambda \in \Lambda} \{\|f(x_1) \circ \lambda - f(x_2)\| \vee \|\lambda - e\|\} \\ &\leq \inf_{\lambda \in \Lambda} \{K\|x_1 \circ \lambda - x_2\| \vee \|\lambda - e\|\} \\ &\leq (K \vee 1) \inf_{\lambda \in \Lambda} \{\|x_1 \circ \lambda - x_2\| \vee \|\lambda - e\|\} \\ &\leq (K \vee 1) d_{J_1}(x_1, x_2) . \end{aligned}$$

For M_1, first note that $f(\Pi(x)) \subseteq \Pi(f(x))$, where $f(\Pi(x)) = \{(f(\tilde{x}), \tilde{t}) : (\tilde{x}, \tilde{t}) \in \Pi(x)\}$. Using (2.13),

$$\begin{aligned} d_{M_1}(f(x_1), f(x_2)) &= \inf_{\substack{(\tilde{x}_i, \tilde{t}_i) \in \Pi(f(x_i)) \\ i=1,2}} \{\|\tilde{x}_1 - \tilde{x}_2\| \vee \|\tilde{t}_1 - \tilde{t}_2\|\} \\ &\leq \inf_{\substack{(\tilde{x}_i, \tilde{t}_i) \in \Pi(x_i) \\ i=1,2}} \{\|f(\tilde{x}_1) - f(\tilde{x}_2)\| \vee \|\tilde{t}_1 - \tilde{t}_2\|\} \\ &\leq \inf_{\substack{(\tilde{x}_i, \tilde{t}_i) \in \Pi(x_i) \\ i=1,2}} \{K\|\tilde{x}_1 - \tilde{x}_2\| \vee \|\tilde{t}_1 - \tilde{t}_2\|\} \\ &\leq (K \vee 1) d_{M_1}(x_1, x_2) . \end{aligned}$$

The continuity follows by similar reasoning. □

Hence, the reflection map in (2.1)–((2.3)) is Lipschitz continuous from (D, M_1) to (D, M_1) and from (D, J_1) to (D, J_1). This continuity is important even with continuous limits to demonstrate measurability, which is needed in the continuous mapping theorem, e.g., in Reiman [25]. (There the limit processes have continuous sample paths, but the processes converging to them do not.) The Lipschitz property also yields rates of convergence for associated weak convergence theorems, applying Whitt [28].

3. An open queueing network with exogenous service interruptions

The model under consideration consists of J service stations which are subject to random service interruptions. A homogeneous customer population arrives from outside the network. Each customer arrives at one of these stations and

requests service. Upon completion of service at one station, a customer may join another station and request service there or may depart from the network.

Let $A_j(t)$ be the cumulative number of customers who arrive at station j from outside the network during the interval $[0, t]$, and let $S_j(t)$ be the cumulative number of customers who are served at station j for the first t units of busy time of that station. (Service times are thus associated with the station instead of the customer.) We call $A \equiv \{A_j, 1 \leqslant j \leqslant J\}$, where $A_j \equiv \{A_j(t), t \geqslant 0\}$ and $S \equiv \{S_j, 1 \leqslant j \leqslant J\}$, where $S_j \equiv \{S_j(t), t \geqslant 0\}$, the *arrival process* and the *service process*, respectively.

The routing of customers is determined by sequences of indicator random variables, $\{\chi_{kj}(l), l = 1, 2, \ldots\}, k, j = 1, 2, \ldots, J$, where $\chi_{kj}(l) = 1$ indicates that the lth departure from station k goes to station j. (It is understood that then $\chi_{km}(l) = 0$ for all $m \neq j$, but this could be changed.) If $\chi_{kj}(l) = 0$ for all j then the $(l+1)^{\text{st}}$ departure from station k leaves the system. For each k and j, let

$$R_{kj}(m) = \sum_{l=1}^{m} \chi_{kj}(l)$$

be the total number among the first m customers who depart station k that go immediately to station j. We call $R \equiv \{R_{kj} : 1 \leqslant k \leqslant J, \quad 1 \leqslant j \leqslant J\}$ with $R_{kj} \equiv \{R_{kj}(m) : m \geqslant 1\}$ the *routing process*.

Let $\{(u_k^j, d_k^j) : k \geqslant 1\}$ be a sequence of random variables, where the u_k^j specifies the duration of the kth up time and d_k^j specifies the duration of the kth down time, both for station j. To be concrete, we assume that all stations start with the beginning of the first up time. Then the epoch beginning the $(l+1)$th up period for station j is

$$T_l^j = \sum_{k=1}^{l} (u_k^j + d_k^j), \quad l \geqslant 1; \quad T_0^j = 0.$$

We assume that $T_l^j \to \infty$ w.p.1. as $l \to \infty$ for each j, so that there are only finitely many up-down cycles in finite time.

Now define an *interruption* (indicator) *process* $I \equiv \{I_j, 1 \leqslant j \leqslant J\}$, where $I_j \equiv \{I_j(t), t \geqslant 0\}$ and $I_j(t) = 1$ indicates that station j is up and $I_j(t) = 0$ indicates that station j is down at time t. Then we have

$$I_j(t) = \begin{cases} 1, & T_l^j \leqslant t < T_l^j + u_{l+1}^j, \\ 0, & T_l^j + u_{l+1}^j \leqslant t < T_{l+1}^j, \end{cases}$$

for $t \geqslant 0, 1 \leqslant j \leqslant J$ and $l \geqslant 0$.

Let $Z_j(0)$ be the initial queue length (number of customers waiting or in service) of station j, $1 \leqslant j \leqslant J$, and let $Z(0) \equiv \{Z_j(0), 1 \leqslant j \leqslant J\}$. The *primitive data* of the model are the arrival process A, the service process S, the routing process R, the interruption process I and the initial queue length $Z(0)$, which we assume are all defined on a common probability space. The dependence structure among these

processes is left for later specification. The processes A, S and I are all assumed to be right-continuous, so that they are elements of D.

The process of primary interest is the *queue length process* $Z \equiv \{Z_j, 1 \leqslant j \leqslant J\}$ with $Z_j \equiv \{Z_j(t), t \geqslant 0\}$, where $Z_j(t)$ is the number of customers at station j at time t. In order to express the queue length process Z in terms of the primitive model data $(A, S, R, I, Z(0))$, we need to introduce some more notation. Let $U_j(t)$ and $D_j(t)$ represent the cumulative up time and down time, respectively, of station j in the interval $[0, t]$ defined by

$$U_j(t) = \int_0^t 1[I_j(s) = 1]\, ds \quad \text{and}$$

$$D_j(t) = t - U_j(t) = \int_0^t 1[I_j(s) = 0]\, ds,$$

where $1A$ is the indicator function of the event A.

Let $B_j(t)$ be the cumulative busy time of station j during the interval $[0, t]$, i.e. the total amount of time during $[0, t]$ station j is serving customers. The *busy-time process* $B \equiv \{B_j, 1 \leqslant j \leqslant J\}$, where $B_j \equiv \{B_j(t), t \geqslant 0\}$, will be expressed in terms of the primitive data below. Then

$$Y_j(t) = U_j(t) - B_j(t) \tag{3.1}$$

is the cumulative idle time of station j (excluding the down time) in the interval $[0, t]$. We can write the queue length process as

$$Z_j(t) = Z_j(0) + A_j(t) + \sum_{k=1}^{J} R_{kj}(S_k(B_k(t))) - S_j(B_j(t)), \quad 1 \leqslant j \leqslant J. \tag{3.2}$$

Note that $S_j(B_j(t))$ gives the actual number of departures from station j during $[0, t]$, so that (3.2) simply expresses the basic conservation of customers at each station: The number of customers present at time t equals the initial number plus the arrivals minus the departures.

To complete the development, we need to specify the busy-time process B. We assume that a work-conserving discipline is used at each station, i.e., the server is always working at full capacity when customers are present at the station and the station is up. Therefore, we must have

$$B_j(t) = \int_0^t 1[Z_j(s) > 0] 1[I_j(s) = 1]\, ds. \tag{3.3}$$

The cumulative-idle-time processes Y can then be written as

$$Y_j(t) = \int_0^t 1[Z_j(s) = 0] 1[I_j(s) = 1]\, ds, \quad t \geqslant 0. \tag{3.4}$$

Similar to the constructive proof of theorem 2.1 in Chen and Mandelbaum [5], we can show that for the given primitive data, there exist a unique pair of processes

(Z, B) with sample paths in D satisfying (3.2) and (3.3). (It suffices to do an induction on the transition epochs of the process (A, S, I).) Clearly, B is continuous in t (actually Lipschitz from (3.3)). Thus, (Z, B) is a well-defined stochastic process on the underlying sample space. Indeed, since (Z, B) over any interval $[0, t]$ depends on $Z(0)$ and finitely many transitions of (A, S, I) over $[0, t]$, (Z, B) is a measurable function of the primitive data. However, the map from $(A, S, R, I, Z(0))$ to Z is in general *not* continuous, even in the case of a one-queue network without disruptions, as shown by example in Whitt [29]. The essential difficulty is that subtraction is not continuous from $D \times D$ to D; see section 4 of [30].

Let λ_j and μ_j be the arrival rate and service rate, respectively, at station j, let P_{jk} be the long-run average fraction of customers who upon departure from station j join station k, and let ν_j be the long-run proportion of time that station j is up. These quantities are formally defined as a consequence of the assumptions (which we now make) that

$$\frac{1}{n} A_j(nt) \to \lambda_j t \quad \text{and} \quad \frac{1}{n} S_j(nt) \to \mu_j t, \tag{3.5}$$

$$\frac{1}{n} R_{kl}([nt]) \to P_{kj} t, \tag{3.6}$$

$$\frac{1}{n} U_j(nt) \to \nu_j t, \tag{3.7}$$

with probability one (w.p.1) as $n \to \infty$ for each k, j and $t > 0$. Since $U_j(t) + D_j(t) = t$, limit (3.7) is equivalent to

$$\frac{1}{n} D_j(nt) \to (1 - \nu_j) t \quad \text{w.p.1} \quad \text{as } n \to \infty. \tag{3.8}$$

Limits (3.5)–(3.8) constitute strong laws of large numbers (SLLNs). (These are equivalent to functional strong laws of large numbers (FSLLNs); see theorem 4 of Glynn and Whitt [14].) For example, limits (3.5)–(3.6) hold if A_j and S_j are renewal processes with finite mean renewal interval for each j, and $R_{kj}(m)$ is a sum of m i.i.d. random variables for each k, j and m, while limits (3.7)–(3.8) hold if $\{(u_k^j, d_k^j) : k \geq 1\}$ is an i.i.d. sequence with finite means. However, these limits also hold more generally.

As a basis for the limit theorems to follow, we present another representation for Z and Y involving the multidimensional reflection map. For each $t \geq 0$, let

$$\begin{aligned} \xi_j(t) = & A_j(t) - \lambda_j t + \sum_{k=1}^{J} [R_{kj}((S_k(B_k(t))) - P_{kj} S_k(B_k(t))] \\ & + \sum_{k=1}^{J} P_{kj} [S_k(B_k(t)) - \mu_k B_k(t)] - [S_j(B_j(t)) - \mu_j B_j(t)], \end{aligned} \tag{3.9}$$

$$\eta_j(t) = (\lambda_j - \mu_j + \sum_{k=1}^{J} \mu_k P_{kj})t + \mu_j D_j(t) - \sum_{k=1}^{J} \mu_k P_{kj} D_k(t)\,, \tag{3.10}$$

$$X_j(t) = Z_j(0) + \xi_j(t) + \eta_j(t), \quad t \geqslant 0\,. \tag{3.11}$$

Then

$$Z_j(t) = X_j(t) + \mu_j Y_j(t) - \sum_{k=1}^{J} \mu_k P_{kj} Y_k(t)\,. \tag{3.12}$$

In vector notation,

$$Z(t) = X(t) + MY(t)\,, \tag{3.13}$$

where

$$M = [I - P^t]\,\mathrm{diag}(\mu)\,. \tag{3.14}$$

Noting that $Z_j(t)$ must be nonnegative, we have

$$\int_0^\infty Z_j(t)\,\mathrm{d}Y_j(t) = \int_0^\infty Z_j(t)1[Z_j(t) = 0]1[I_j(t) = 1]\,\mathrm{d}t = 0\,. \tag{3.15}$$

Therefore, the queue length process Z and the cumulative idle time process Y are related to X in (3.11) via the reflection mapping defined in section 2; i.e., $Q = P^t$, $\psi(X) = Z$ and $\phi(X) = \mathrm{diag}(\mu)\,Y$.

Remark (3.1)

From section 2, it follows that (Y, Z) is a continuous function of X, but after (3.3) we noted that Z is not a continuous function of the primitive data $(A, S, R, I, Z(0))$. The explanation is that X is not a continuous function of $(A, S, R, I, Z(0))$. □

To conclude this section, we provide some definitions which classify the stations of the network. The details and their validations can be found in Chen and Mandelbaum [5]. (From the definition of the routing process, the matrix $P \equiv (P_{jk})$ is a substochastic matrix.) The unique solution of

$$x = \lambda + P^t(x \wedge \mu)\,, \tag{3.16}$$

denoted by λ^e is called the *effective arrival rate*, and

$$\rho_j \equiv \lambda_j^e / \mu_j \tag{3.17}$$

is called the *traffic intensity* of station j. A station j is a *non-bottleneck station* if $\rho_j < \nu_j$, and a *bottleneck station* if $\rho_j \geqslant \nu_j$. In the bottleneck case, we call station j a *balanced bottleneck* if $\rho_j = \nu_j$, and otherwise, a *strict bottleneck*. The network is referred to as a *balanced network* if $\rho_j = \nu_j$ for all j.

We remark that the bottleneck definitions only depend on the parameter four-

tuple (λ, μ, P, ν). So we will also say that station j is a non-bottleneck of network (λ, μ, P, ν), and that (λ, μ, P, ν) is a balanced network if $\rho_j = \nu_j$, for all j, $j = 1, \ldots, J$, with ρ_j defined through λ, μ and P.

4. Diffusion limit with jumps for long up and down times

Consider a sequence of open networks with exogenous service interruptions, indexed by $n = 1, 2, \ldots$. We add a superscript n to all processes, variables and parameters, associated with the nth network. In this section, we consider the convergence of the scaled queue length process $(1/\sqrt{n})Z^n(nt)$ under the condition that the up times are of order n, while the down times are of order $\sqrt{n}$, in addition to the usual heavy-traffic assumptions, i.e., $(1 - \rho_j^n)$ is of order $1/\sqrt{n}$. This section thus contains the network generalization of the heavy-traffic limit theorem in section 3 of Kella and Whitt [22].

(A) Assumptions on the primitive data

First, for simplicity, we assume that the routing structures are the same for all networks, i.e., $R^n = R$ for all $n \geq 1$. Then we assume that as $n \to \infty$, *jointly*

$$\frac{1}{\sqrt{n}} Z^n(0) \Rightarrow \hat{Z}(0) \quad \text{in } \mathbb{R}^j, \tag{4.1}$$

$$\frac{1}{\sqrt{n}} [A^n(nt) - \lambda^n nt] \Rightarrow \hat{A}(t) \quad \text{in } D^J, \tag{4.2}$$

$$\frac{1}{\sqrt{n}} [S^n(nt) - \mu^n nt] \Rightarrow \hat{S}(t) \quad \text{in } D^J, \tag{4.3}$$

$$\frac{1}{\sqrt{n}} [R([nt]) - Pnt] \Rightarrow \hat{R}(t) \quad \text{in } D^{J^2}, \tag{4.4}$$

$$\left\{ \left(\frac{u_k^{j,n}}{n}, \frac{d_k^{j,n}}{\sqrt{n}} \right), 1 \leq j \leq J, k \geq 1 \right\} \Rightarrow \left\{ (u_k^j, d_k^j), 1 \leq j \leq J, k \geq 1 \right\} \quad \text{in } (\mathbb{R}^{2J})^\infty, \tag{4.5}$$

$$\sqrt{n}(\lambda^n - \lambda) \to c_\lambda, \quad \lambda \geq 0, \quad \text{and} \tag{4.6}$$

$$\sqrt{n}(\mu^n - \mu) \to c_\mu, \quad \mu > 0, \tag{4.7}$$

where $\sum_{k=1}^m u_k^j \to \infty$ w.p.1. as $m \to \infty$ for each j and the limit processes $\hat{A}$, $\hat{S}$ and $\hat{R}$ have continuous sample paths w.p.1. (This continuity can be relaxed, i.e., replaced by other assumptions, but in most applications the processes $\hat{A}$, $\hat{S}$ and $\hat{R}$ are Brownian motions, which do have continuous sample paths.) Note that we require joint convergence of (4.1)–(4.5). The standard sufficient condition is mutual indepen-

dence; see theorem 3.2 of Billingsley [2]. However, in some cases of interest it is important to relax the independence; e.g., see Fendick, Saksena and Whitt [11].

In this section, we focus on the case of asymptotically balanced open networks. First, note that (3.7) and (4.5) imply that $\nu_j = 1$ for all j; i.e., asymptotically as $n \to \infty$ the proportion of time each station is up is 1. (Nevertheless, as in [22], the down times have a significant impact on the limit by introducing jumps.) Thus, to have the limiting network (λ, μ, P, ν) specified by (4.4), (4.5), (4.6) and (4.7) balanced, we assume that

$$\lambda = [I - P^t]\mu, \tag{4.8}$$

where the matrix P is substochastic with $P^k \to 0$ as $k \to \infty$.

(B) The cumulative down time limit

In preparation for the main theorem, we establish convergence for the sequence of cumulative down-time processes. For this purpose, let $\hat{N}_j(t)$ be the counting process associated with u_k^j in (4.5), i.e.,

$$\hat{N}_j(t) = \sup\left\{m \geqslant 0 : \sum_{k=1}^{m} u_k^j \leqslant t\right\}, \quad t \geqslant 0, \quad \text{and}$$

$$\hat{D}_j(t) = \sum_{k=1}^{\hat{N}_j(t)} d_k^j, \quad t \geqslant 0. \tag{4.9}$$

By the assumption about u_k^j in section 4(A), $P(\hat{N}_j(t) < \infty) = 1$ for all j and t. Let $\hat{D}^n$ be the normalized cumulative down-time process in model n, defined by

$$\hat{D}_j^n(t) = \frac{1}{\sqrt{n}} D_j^n(nt), \quad t \geqslant 0, 1 \leqslant j \leqslant J. \tag{4.10}$$

Let $\hat{D}^n \equiv \{\hat{D}_j^n, 1 \leqslant j \leqslant J\}$ and $\hat{D} \equiv \{\hat{D}_j, 1 \leqslant j \leqslant J\}$ be the associated vector processes in D^J. Let Disc(x) be the set of discontinuity points of x in $[0, T]$.

LEMMA 4.1

If (4.5) holds and

$$P\left(\bigcup_{i=1}^{J}\bigcup_{\substack{j=1 \\ j \neq i}}^{J}(\text{Disc}(\hat{D}_i) \cap \text{Disc}(\hat{D}_j)) = \phi\right) = 1, \tag{4.11}$$

then

$$\hat{D}^n \Rightarrow \hat{D} \quad \text{in } D((0, \infty), \mathbb{R}^J, M_1).$$

Proof

First apply the Skorohod representation theorem to replace the weak convergence assumed in (4.5) by w.p.1. convergence. Then it is elementary that $\hat{D}_j^n(t) \to \hat{D}_j(t)$ w.p.1. for each t that is not a discontinuity point of $\hat{D}_j(t)$. Since $\hat{D}_j^n(t)$ and $\hat{D}_j(t)$ are nondecreasing, this implies convergence in $D((0,\infty),\mathbb{R},M_1)$; see the remark after theorem 7.1 on p. 82 of Whitt [30]. This M_1 convergence holds for any closed time interval $[a,b]$ provided that a and b are not points of discontinuity of $\hat{D}$ with positive probability. Since 0 could be a point of discontinuity of $\hat{D}$ with positive probability, we work with the open time interval $(0,\infty)$. However, so far this argument only takes care of one coordinate at a time; i.e., this argument yields the convergence in the product space $D((0,\infty),\mathbb{R},M_1)^J$ whereas we want to establish the more difficult convergence in $D((0,\infty),\mathbb{R}^J,M_1)$, which involves only a single parametric representation. For this purpose, note that since (4.11) holds with $P(\hat{N}_j(t)<\infty)=1$ for each j and t, the limit process $\hat{D}$ has only finitely many discontinuities in $[0,T]$ w.p.1. Let $\Delta = \mathrm{Disc}(\hat{D}) \cap \{0\} = \{t_i : 1 \leqslant i \leqslant m\}$ for one sample point. Then we can establish convergence $\hat{D}^n \to \hat{D}$ in $D([s_i,s_{i+1}],\mathbb{R}^J,m_1)$ for each i, $1 \leqslant i \leqslant m-1$, where $s_i = (t_i + t_{i+1})/2$ by constructing the appropriate parametric representation needed for the one coordinate that has a discontinuity in (s_i,s_{i+1}). Since $\hat{D}_j$ has no discontinuities in (s_i,s_{i+1}) for the other j, we have $\hat{D}_j^n \to \hat{D}_j$ uniformly on $[s_i,s_{i+1}]$ for these other j and thus also in the M_1 topology using the same parametric representation as used for the coordinate with the discontinuity. Finally, we piece together the parametric representations to obtain one parametric representation for the interval $[s_1,s_{m-1}]$. In particular, for the subinterval $[s_i,s_{i+1}]$ consider the parametric representations mapping $[(i-1)/(m-1), i/(m-1)]$ onto $\Gamma_{\hat{D}^n}$ and $\Gamma_{\hat{D}}$, $1 \leqslant i \leqslant m-1$. This yields standard parametric representations mapping $[0,1]$ onto $\Gamma_{\hat{D}^n}$ and $\Gamma_{\hat{D}}$ associated with time interval $[s_1,s_{m-1}]$. (Necessarily the parametric representations of $\hat{D}$ map $(i-1)/(m-1)$ and $i/(m-1)$ into $(\hat{D}(s_i),s_i)$ and $(\hat{D}(s_{i+1}),s_{i+1})$, respectively, so the endpoints match. Since this argument works for any $T>0$ and any s_1', $0<s_1'<s_1$ when we replace s_1 by s_1', we have established convergence in $D((0,\infty),\mathbb{R}^J,M_1)$. Since this holds for each sample point after applying the Skorohod representation theorem, we have the claimed weak convergence for the original processes.

Remarks

(4.1) The standard sufficient condition for the discontinuity condition (4.11) is to have the sequences $\{u_k^j : k \geqslant 1\}$ be mutually independent with $\sum_{k=1}^m u_k^j$ having a continuous cdf for each $m \geqslant 1$ and $j \geqslant 2$. This does not require that the random variables within each sequence be mutually independent. If the random variables u_k^j, $k \geqslant 1$, are mutually independent, then $\sum_{k=1}^m u_k^j$ will have a continuous cdf for all m if u_1^j does.

(4.2) To see that we do not have convergence in the standard J_1 topology in lemma 4.1, note that $\hat{D}^n$ is continuous for each n, while $\hat{D}$ is not. (In the J_1 topology,

the maximum jump functional is continuous.) To see that we need not have convergence for the time interval $[0,\infty)$, suppose that $P(u_1^{1,n} = \sqrt{n}) = P(d_1^{1,n} = \sqrt{n}) = 1$ so that $P(u_1^1 = 0) = P(d_1^1 = 1) = 1$. Then, from (4.5), we have $P(\hat{D}_1^n(0) = 0) = 1$ but $P(\hat{D}_1(0) > 0) > 0$.

THE MAIN LIMIT THEOREM

To state the main limit theorem, let

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

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

$$\hat{B}^n(t) \equiv \frac{1}{\sqrt{n}} [B^n(nt) - nt], \quad t \geq 0, \tag{4.12}$$

be the scaled queue-length, cumulative-idle-time and cumulative-busy-time processes, respectively.

THEOREM 4.1

If (4.11) and the assumptions of section 4(A) hold, then

$$(\hat{Z}^n, \hat{B}^n, \hat{Y}^n, \hat{D}^n \Rightarrow (\hat{Z}, \hat{B}, \hat{Y}, \hat{D}) \text{ in } D((0,\infty), \mathbb{R}^{4J}, M_1),$$

where

$$\hat{Z} = \phi(\hat{X}) \quad \text{and} \quad \hat{Y} = \text{diag}(\mu^{-1})\psi(\hat{X}). \tag{4.13}$$

$$\hat{X}(t) = \hat{Z}(0) + \hat{\xi}(t) + \hat{\eta}(t), \tag{4.14}$$

$$\hat{\xi}_j(t) = \hat{A}_j(t) + \sum_{k=1}^{J} [\hat{R}_{kj}(\mu_k t) + P_{kj}\hat{S}_k(t)] - \hat{S}_j(t) \quad \text{for each } j, \tag{4.15}$$

$$\hat{\eta}(t) = (c_\lambda - [I - P^t]c_\mu)t + [I - P^t]\,\text{diag}(\mu)\hat{D}(t), \tag{4.16}$$

$$\hat{B}(t) = -\hat{D}(t) - \hat{Y}(t) \tag{4.17}$$

with $\hat{D}$ in (4.9), and (ψ, ϕ) the reflection map in section 2.

Remarks

(4.3) Let $W_j^n(t)$ represent the workload in remaining service time at node j at time t in model n. Let $\hat{W}^n \equiv W^n(nt)/\sqrt{n}$ be the associated normalized vector workload process. By essentially the same argument as in sections 5.5 and 6.8 of Chen and Mandelbaum [6], we can conclude that $\hat{W}^n \Rightarrow \hat{W}$ as $n \to \infty$ jointly with the processes in theorem 4.1 in $D((0,\infty), \mathbb{R}^{5J}, M_1)$, where $\hat{W} = \text{diag}(\mu^{-1})\hat{Z}$. This result can be regarded as a diffusion analog of Little's law.

(4.4) In theorem 4.1 we made no assumptions of independence for the arrival, service and routing processes. In many applications (e.g., when A_j^n and S_j^n are

renewal processes and the routing R is Markovian), the limit processes $\hat{A}$, $\hat{S}$ and $\hat{R}$ in (4.2)–(4.4) will be Brownian motions, in which case $\hat{\xi}$ in (4.15) will be as well.

Proof of theorem 4.1

By assumption in section 4(A), the limits in (4.1)–(4.5) hold jointly. By the Skorohod representation theorem, there exists a probability space on which versions of all stochastic processes and random variables in (4.1)–(4.5) are defined with convergence holding almost surely, where the topology on D is uniform convergence on compact subintervals (u.o.c.). From these new versions of primitive data, we can construct new versions of the stochastic processes of interest, namely, $(\hat{Z}^n, \hat{B}^n, \hat{Y}^n, \hat{D}^n)$ as indicated in section 3. Then the proof amounts to showing that the new version of $(\hat{Z}^n, \hat{B}^n, \hat{Y}^n, \hat{D}^n)$ converges almost surely to the associated limit $(\hat{Z}, \hat{B}, \hat{Y}, \hat{D})$ in the M_1 topology.

For simplicity, we use the same notation for the new versions of all processes and random variables. Specifically, we will prove

$$(\hat{Z}^n, \hat{B}^n, \hat{Y}^n, \hat{D}^n) \to (\hat{Z}, \hat{B}, \hat{Y}, \hat{D}) \quad \text{in } D((-\infty, \infty), \mathbb{R}^{4J}, M_1)$$

(almost surely) as $n \to \infty$ under assumptions (4.6)–(4.8) and

$$\frac{1}{\sqrt{n}} Z^n(0) \to \hat{Z}(0), \quad \text{u.o.c.}, \tag{4.18}$$

$$\frac{1}{\sqrt{n}} [A^n(nt) - \lambda^n nt] \to \hat{A}(t), \quad \text{u.o.c.}, \tag{4.19}$$

$$\frac{1}{\sqrt{n}} [S^n(nt) - \mu^n nt] \to \hat{S}(t), \quad \text{u.o.c.}, \tag{4.20}$$

$$\frac{1}{\sqrt{n}} [R_{kj}([nt]) - P_{kj} nt] \to \hat{R}_{kj}(t), \quad \text{u.o.c.}, \tag{4.21}$$

$$\left\{ \left[\frac{u_k^{j,n}}{n}, \frac{d_k^{j,n}}{\sqrt{n}} \right], 1 \leqslant j \leqslant J, k \geqslant 1 \right\} \to \left\{ (u_k^j, d_k^j), 1 \leqslant j \leqslant J, k \geqslant 1 \right\}, \tag{4.22}$$

(almost surely) as $n \to \infty$.

We use the following lemma, which is proved at the end of this section.

LEMMA 4.2

Under the assumptions above,

$$\frac{1}{\sqrt{n}} B_j^n(nt) \to t, \quad \text{u.o.c.}, \quad \text{as } n \to \infty, 1 \leqslant j \leqslant J.$$

From (3.9), we see that

$$\begin{aligned}\frac{1}{\sqrt{n}}\xi_j^n(nt) = &\frac{1}{\sqrt{n}}[A_j^n(nt) - \lambda_j^n nt] \\ &+ \sum_{k=1}^{J}\frac{1}{\sqrt{n}}\left[R_{kj}\left(n\frac{1}{n}S_k^n\left(n\frac{B_k^n(nt)}{n}B_k^n(nt)\right)\right) - P_{kj}n\frac{1}{n}S_k^n\left(n\frac{1}{n}B_k^n(nt)\right)\right] \\ &+ \sum_{k=1}^{J}\frac{1}{\sqrt{n}}\left[S_k^n\left(n\frac{B_k^n(nt)}{n}\right) - \mu_k^n n\frac{B_k^n(nt)}{n}\right]P_{kj} \\ &- \frac{1}{\sqrt{n}}\left[S_j^n\left(n\frac{B_j^n(nt)}{n}\right) - \mu_j^n n\frac{B_j^n(nt)}{n}\right]. \end{aligned} \tag{4.23}$$

Using lemma 4.2 and the composition map (a deterministic version of the random-time-change theorem; see sections 3 and 5 of Whitt [30] and 5.1.D of Chen and Mandelbaum [6]), we see that as $n \to \infty$

$$\frac{1}{\sqrt{n}}\xi_j^n(nt) \to \hat{\xi}_j(t), \quad \text{u.o.c.}, \tag{4.24}$$

where $\hat{\xi}_j$ is as defined in (4.15), $j = 1, 2, \ldots, J$, and is continuous.

By lemma 4.1,

$$\frac{1}{\sqrt{n}}D^n(nt) \to \hat{D}(t) \quad \text{in } D((0,\infty), \mathbb{R}^J, M_1) \quad \text{as } n \to \infty. \tag{4.25}$$

From (3.10), we see that

$$\begin{aligned}\frac{1}{\sqrt{n}}\eta^n(nt) &= \sqrt{n}(\lambda^n - [I - P^t]\mu^n) + [I - P^t]\,\mathrm{diag}(\mu^n)\frac{1}{\sqrt{n}}D^n(nt) \\ &= (\sqrt{n}(\lambda^n - \lambda) - [I - P^t]\sqrt{n}(\mu^n - \mu))t + [I - P^t]\,\mathrm{diag}(\mu^n)\frac{1}{\sqrt{n}}D^n(nt). \end{aligned} \tag{4.26}$$

From (4.6), (4.7), (4.8), (4.25) and (4.26), we obtain

$$\frac{1}{\sqrt{n}}\eta^n(nt) \to \hat{\eta}(t) \quad \text{in } D((0,\infty), \mathbb{R}^J, M_1) \tag{4.27}$$

as $n \to \infty$, for $\hat{\eta}$ in (4.16).

From (3.11), we obtain

$$\hat{X}^n(t) \equiv \frac{1}{\sqrt{n}}X^n(nt) = \frac{1}{\sqrt{n}}Z^n(0) + \frac{1}{\sqrt{n}}\xi^n(nt) + \frac{1}{\sqrt{n}}\eta^n(nt), \quad t \geqslant 0.$$

By (4.18), (4.24), (4.27) plus the measurability and continuity w.p.1 with respect to the limit (ξ, η) of addition (recall that ξ is continuous; see theorem 4.1 of Whitt [30], we have

$$\frac{1}{\sqrt{n}}X^n(nt) \to \hat{X}(t) \quad \text{in } D((0,\infty), \mathbb{R}^J, M_1) \tag{4.28}$$

as $n \to \infty$ for $\hat{X}$ in (4.14). Moreover, by the assumed joint convergence in (4.1)–(4.5), we can actually have the joint convergence

$$\left(\frac{1}{\sqrt{n}} X^n(nt), \frac{1}{\sqrt{n}} D^n(nt)\right) \to (\hat{X}(t), \hat{D}(t)) \quad \text{in } D((0,\infty), \mathbb{R}^{2J}, M_1) \tag{4.29}$$

as $n \to \infty$. The parametric representations used for $\hat{D}^n \to \hat{D}$ can be used with all other processes.) Then

$$\hat{Y}^n(t) \equiv \frac{1}{\sqrt{n}} Y^n(nt) = \text{diag}(\mu^{-1}) \psi\left(\frac{1}{\sqrt{n}} X^n(n\cdot)\right)(t), \tag{4.30}$$

$$\hat{Z}^n(t) \equiv \frac{1}{\sqrt{n}} Z^n(nt) = \phi\left(\frac{1}{\sqrt{n}} X^n(n\cdot)\right)(t) \quad \text{and} \tag{4.31}$$

$$\begin{aligned} \hat{B}^n(t) &= \frac{1}{\sqrt{n}} [U^n(nt) - nt] - \frac{1}{\sqrt{n}} Y^n(nt) \\ &= -\hat{D}^n(t) - \hat{Y}^n(t). \end{aligned} \tag{4.32}$$

by the reflection map representation in section 3. Hence, we have

$$(\hat{X}^n, \hat{Z}^n, \hat{B}^n, \hat{Y}^n, \hat{D}^n) \to (\hat{X}, \hat{Z}, \hat{B}, \hat{Y}, \hat{D}) \quad \text{in } D((0,\infty), \mathbb{R}^{5J}, M_1)$$

as $n \to \infty$ by continuity of the oblique reflection mapping using (4.29)–(4.32) and section 2. □

Proof of lemma 4.2

Since

$$\begin{aligned} \left|\frac{1}{n} B^n(nt) - \frac{1}{n} B^n(ns)\right| &= \frac{1}{n} \left| \int_{ns}^{nt} 1[Z_k^n(u) > 0, L_k^n(u) = 1] \, du \right| \\ &= \left| \int_s^t 1[Z_k^n(nu) > 0, L_k^n(nu) = 1] \, du \right| \leqslant |t - s|, \end{aligned} \tag{4.33}$$

the sequence $\{n^{-1} B^n(nt), n \geqslant 1\}$ is uniformly Lipschitz. By Ascoli's theorem, any subsequences of $\{n^{-1} B^n, n \geqslant 1\}$ has an u.o.c. convergent subsequence. So we only need to prove any u.o.c. limit of any subsequence is the same and given by t. Without loss of generality (for ease of notation), assume

$$\frac{1}{n} B^n(nt) \to \bar{B}(t), \quad \text{u.o.c.} \quad \text{as } n \to \infty. \tag{4.34}$$

We will show that $\bar{B}(t) = t$. From (4.23), (4.19)–(4.21) and (4.33) it implies that

$$\frac{1}{n} \xi^n(nt) \to 0, \quad \text{u.o.c.,} \quad \text{as } n \to \infty. \tag{4.35}$$

Note that

$$\frac{1}{n}\eta^n(t) = \theta^n t + [I - P']\,\mathrm{diag}(\mu^n)\frac{D^n(t)}{n},$$

where $\theta^n = \lambda^n - (I - P')\mu^n \to 0$ as $n \to \infty$. Hence,

$$\frac{1}{n}\eta^n(nt) \to 0, \quad \text{u.o.c.}, \quad \text{as } n \to \infty, \tag{4.36}$$

because $\mu^n \to \mu$ by (4.7) and

$$\frac{1}{n}D^n(nt) \to 0, \quad \text{u.o.c.}, \quad \text{as } n \to \infty \tag{4.37}$$

as a consequence of lemma 4.1. Combining (4.18), (4.35) and (4.36), we have

$$\frac{1}{n}X^n(nt) \to \bar{X}(t) \equiv 0, \quad \text{u.o.c.}, \quad \text{as } n \to \infty. \tag{4.38}$$

From the continuity and the uniqueness of the reflection mapping, we have

$$\frac{1}{n}Y^n(nt) \equiv \phi\left(\frac{1}{n}X^n(n\cdot)\right)(t) \to \bar{Y}(t) \equiv \phi(\bar{X})(t) \equiv 0, \quad \text{u.o.c.}, \quad \text{as } n \to \infty. \tag{4.39}$$

Since

$$B^n(t) = U^n(t) - Y^n(t) = t - D^n(t) - Y^n(t), \tag{4.40}$$

we prove $\bar{B}(t) = t$ by using (4.37) and (4.39). □

5. Diffusion limit with fixed up and down times

We now establish the diffusion limit corresponding to the up and down times $(u_k^{j,n}, d_k^{j,n})$ being independent of n instead of growing with n as specified by (4.5). In particular, we keep all the assumptions of section 4(A) except (4.5) and assume instead that

$$\frac{1}{\sqrt{n}}[D^n(nt) - (1 - \nu^n)nt] \to \hat{D}(t) \quad \text{in } D^J \tag{5.1}$$

jointly with (4.1)–(4.4), where $\hat{D}(t)$ has continuous sample paths w.p.1 and

$$\sqrt{n}(\nu^n - \nu) \to c_\nu, \quad 0 \leqslant \nu \leqslant 1, \tag{5.2}$$

as $n \to \infty$. Note that (5.1) and (5.2) imply that

$$\frac{1}{\sqrt{n}}[D^n(nt) - (1 - \nu)nt] \to \hat{D}(t) - c_\nu t \quad \text{in } D^J \tag{5.3}$$

by virtue of theorems 4.4 and 5.1. of Billingsley [2].

A sufficient condition for (5.3) with $c_\nu = 0$ is for the variables $(u_k^{j,n}, d_k^{j,n})$ to be independent of n, mutually independent and have a common distribution as $k \geqslant 1$ for each j. Then the up-down processes are J independent alternating renewal pro-

cesses. (Of course, (5.3) holds with $c_{\nu j} = 0$ for all j under many other conditions too.)

Before stating the general result under (5.1) and (5.2), we describe the limit (5.3) in the alternating-renewal-process case. In this alternating-renewal-process context, suppose that $Eu_1^j = u_j$, $\mathrm{Var}\, u_1^j = \sigma_{uj}^2$, $0 < \sigma_{uj}^2 < \infty$, $Ed_1^j = d_j$ and $\mathrm{Var}\, d_1^j = \sigma_{dj}^2$, $0 < \sigma_{dj}^2 < \infty$. Then (5.3) is valid with $\nu_j = u_j/(u_j + d_j)$ and $c_{\nu j} = 0$ for all j, and $\hat{D} \equiv (\hat{D}_1, \ldots, \hat{D}_J)$ being composed of J independent zero-mean one-dimensional Brownian motions with diffusion (variance) constants specified below.

THEOREM 5.1

In the alternating-renewal-process context above, (5.3) is valid with $\nu_j = u_j/(u_j + d_j)$ and $c_{\nu j} = 0$ for all j, and the variance constant for the jth coordinate Brownian motion is

$$\sigma_j^2 = \lambda_j(1 - \lambda_j d_j)^2 c_{dj}^2 + \lambda_j^3 d_j^2 \sigma_{uj}^2 , \tag{5.4}$$

where $\lambda_j = 1/(d_j + u_j)$.

Proof

We use the continuous mapping theorem with the functions in Whitt [30]. Similar results are established in section 3 of Glynn and Whitt [13]. In particular, let

$$S_n^u = \sum_{k=1}^{n} u_k^j, \quad S_n^d = \sum_{k=1}^{n} d_k^j \quad \text{and} \quad S_n = S_n^u + S_n^d, \quad n \geqslant 1 . \tag{5.5}$$

By the independence and moment assumptions above

$$\frac{1}{\sqrt{n}}[S_{[nt]} - (u_j + d_j)nt, S_{[nt]}^d - d_n nt] \Rightarrow (\sigma_{uj} B_u(t) + \sigma_{dj} B_d(t), \sigma_{dj} B_d(t)) \quad \text{in } D([0,\infty), \mathbb{R}^2, J_1) , \tag{5.6}$$

where B_u and B_d are independent standard (mean 0, variance 1) Brownian motions. Let $N(t)$ be the renewal counting process associated with the partial sums S_u and let $\lambda_j = 1/(d_j + u_j)$. By the inverse map in section 7 of [30],

$$\frac{1}{\sqrt{n}}[N(nt) - \lambda_j nt, S_{[nt]}^d - d_j nt] \Rightarrow [-\lambda_j \sigma_{uj} B_u(\lambda_j t) - \lambda_j \sigma_{dj} B_d(\lambda_j t), \sigma_{dj} B_d(t)] \quad \text{in } D([0,\infty), \mathbb{R}^2, J_1) \tag{5.7}$$

By composition plus translation in section 5 of [30],

$$\frac{1}{\sqrt{n}}[S_{N(nt)}^d - \lambda_j d_j nt] \Rightarrow \sigma_{dj} B_d(\lambda_j t) - d_j \lambda_j \sigma_{uj} B_u(\lambda_j t) - d_j \lambda_j \sigma_{dj} B_d(\lambda_j t) \quad \text{in } D([0,\infty), \mathbb{R}, J_1) . \tag{5.8}$$

By theorem 4.1 of [2], $S_{N(t)}^d$ has the same FCLT behavior as $D_j(t)$. Hence, (5.3) is valid with $(1 - \nu_j) = \lambda_j d_j = d_j/(u_j + d_j)$, $c_{\nu j} = 0$ and

$$\hat{D}_j(t) = (1-\lambda_j d_j)\sigma_{dj} B_d(\lambda_j t) - \lambda_j d_j \sigma_{uj} B_u(\lambda_j t)$$
$$\stackrel{\mathrm{d}}{=} (\lambda_j(1-\lambda_j d_j)^2 \sigma_{dj}^2 + \lambda_j^3 d_j^2 \sigma_{uj}^2)^{1/2} B(t), \tag{5.9}$$

where $\stackrel{\mathrm{d}}{=}$ denotes equal in distribution and B is a standard Brownian motion. □

Now to have an asymptotically balanced network, instead of (4.8), we assume that

$$\text{Rate in to } j \equiv \lambda_j + \sum_{k=1}^{J} P_{kj}\mu_k\nu_k = \mu_j\nu_j \equiv \text{Rate out of } j \quad \text{for each } j. \tag{5.10}$$

Let $\hat{Z}^n$ and $\hat{Y}^n$ be defined as in (4.12), but redefine $\hat{B}^n$ as

$$\hat{B}^n(t) = \frac{1}{\sqrt{n}}[B^n(nt) - \nu^n nt], \quad t \geqslant 0. \tag{5.11}$$

Paralleling lemma 4.2, we use the following lemma, which we prove at the end of this section.

LEMMA 5.1

As $n \to \infty$,

$$\frac{B^n(nt)}{n} \to \nu t \quad \text{w.p.1} \quad \text{in } D([0,\infty), \mathbb{R}^J, J_1).$$

Here is our main result in this section. We omit the proof because it is similar to the proof of theorem 4.1.

THEOREM 5.2

If the assumptions of section 4.1 hold with (5.1), (5.2) and (5.10) instead of (4.5), then

$$(\hat{Z}^n, \hat{B}^n, \hat{Y}^n, \hat{D}^n) \Rightarrow (\hat{Z}, \hat{B}, \hat{Y}, \hat{D}) \quad \text{in } D([0,\infty), \mathbb{R}^{4J}, J_1),$$

where $\hat{Z}$, $\hat{Y}$, $\hat{X}$, $\hat{B}$ and $\hat{D}$ are defined as in (4.13), (4.14), (4.17) and (5.1),

$$\hat{\xi}_j(t) = \hat{A}_j(t) + \sum_{k=1}^{J}[\hat{R}_{kj}(\mu_k\nu_k t) + P_{kj}\hat{S}_k(\nu_k t)] - \hat{S}_j(\nu_j t) \quad \text{for each } j, \tag{5.12}$$

and

$$\hat{\eta}_j(t) = \left[c_{\lambda j} - (\mu_j c_{\nu j} + \nu_j c_{\mu j}) + \sum_{k=1}^{J} P_{kj}(\mu_k c_{\nu k} + \nu_k c_{\mu k})\right] t$$
$$+ \mu_j \hat{D}_j(t) - \sum_{k=1}^{J} \mu_k P_{kj} \hat{D}_k(t) \quad \text{for each } j. \tag{5.13}$$

Remarks (5.1)

In the standard case $\hat{A}, \hat{R}, \hat{S}$ and $\hat{D}$ are independent zero-mean Brownian motions, in which case $\hat{X}$ is a Brownian motion and $\hat{Z}$ is regulated Brownian motion (RBM) as in Harrison and Reiman [17]. The service interruptions affect the limit though the asymptotic up-time parameters ν_j appearing as deterministic time changes in some terms in $\hat{\xi}_j$ in (5.12) and by the additional term $\mu_j \hat{D}_j(t) - \sum_{k=1}^{J} \mu_k P_{kj} \hat{D}_k(t)$ in $\hat{\eta}_j(t)$ in (5.13). If $\hat{D}$ is a zero-mean Brownian motion with covariance matrix C, then $(I - P^t)\,\mathrm{diag}(\mu)\hat{D}$ is a zero-mean Brownian motion with covariance matrix $(I - P^t)\,\mathrm{diag}(\mu)C\,\mathrm{diag}(\mu)(I - P)$; see p. 82 of Feller [10].

Proof of lemma 5.1

To establish lemma 5.1, proceed as in the proof of lemma 4.2, but note that now $\bar{B}(t) = \nu t$. From (5.3), we see that

$$\frac{1}{n} D^n(nt) \to (1 - \nu)t, \quad \text{u.o.c.} \tag{5.14}$$

instead of (4.37). Then apply (4.38)–(4.40) to obtain the desired result. □

6. Concluding remarks

We have established heavy-traffic limits for two cases of open queueing networks with exogenous service disruptions. The "standard" case of fixed up and down times was treated in section 5; it leads to an RBM limit so that it falls within the domain of much existing theory, e.g., Harrison and Reiman [17], Reiman [25], Harrison and Williams [18,19] and Chen and Mandelbaum [6].

What we regard as more interesting is the case of long up and down times (of order n and $\sqrt{n}$, respectively) which was treated in section 4; it is the network generalization of the heavy-traffic limit in section 3 of Kella and Whitt [22], which was briefly discussed in remark 5.6 in [22]. Unlike the one-dimensional case in [22], we have yet to obtain useful descriptions of the stationary distribution of the limiting multidimensional "jump-diffusion" process.

Unlike [22], we have not discussed the case of service interruptions or vacations occurring whenever a station becomes empty. Heavy-traffic limits for the case also follow, just as in [22], by treating the subintervals between successive vacations separately, but we do not present the details. As for the case of exogenous service interruptions treated explicitly here in section 4, it remains to describe the stationary distribution of the limiting multidimensional "jump-diffusion" process.

We have also not discussed closed networks, but irreducible closed networks can be treated similarly, drawing on Chen and Mandelbaum [4–6].

Chen and Mandelbaum [7] have recently established strong approximation for open queueing networks. These results can be extended to the models considered

here, thus providing rates of convergence for the limit theorems established here; see Horvath [20] and Glynn and Whitt [15] for related work.

References

[1] S. Asmussen, The heavy traffic limit of a class of Markovian queueing models, Oper. Res. Lett. 6 (1988) 301–306.
[2] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).
[3] D.Y. Burman, Approximations for a service system with interruptions, AT&T Bell Laboratories, Murray Hill, NJ (1987).
[4] H. Chen and A. Mandelbaum, Leontief systems, RBV's and RBM's, in: *Proc. Imperial College Workshop on Applied Stochastic Processes*, eds. M.H.A. Davis and R.J. Elliot (Gordon and Breach, London, 1991).
[5] H. Chen and A. Mandelbaum, Discrete flow networks: bottleneck analysis and fluid approximation, Math. Oper. Res. 16 (1991) 408–446.
[6] H. Chen and A. Mandelbaum, Discrete flow networks: diffusion approximations and bottlenecks, Ann. Prob., 19 (1991) 1463–1519.
[7] H. Chen and A. Mandelbaum, Strong approximations for open queueing networks, in preparation.
[8] E.G. Coffman, Jr., and M.I. Reiman, Diffusion approximations for computer communication systems, in: *Mathematical Computer Performance and Reliability*, eds. G. Iazeolla, P.J. Courtois and A. Hordijk (Elsevier, Amsterdam, 1984) pp. 33-53.
[9] P. Dupuis and H. Ishii, On when the solution to the Skorohod problem is Lipschitz continuous, with applications, Department of Mathematics and Statistics, University of Massachusetts, Amherst (1989).
[10] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, 2nd ed. (Wiley, New York, 1971).
[11] K.W. Fendick, V.R. Saksena and W. Whitt, Dependence in packet queues, IEEE Trans. Commun. 37 (1989) 1173–1183.
[12] M.J. Fischer, An approximation for queueing systems with interruptions, Manag. Sci. 24 (1977) 338–344.
[13] P.W. Glynn and W. Whitt, A central-limit- theorem version of $L = \lambda W$, Queueing Systems 2 (1986) 191–215.
[14] P.W. Glynn and W. Whitt, Ordinary CLT and WLLN versions of $L = \lambda W$, Math. Oper. Res. 13 (1988) 674–692.
[15] P.W. Glynn and W. Whitt, Departures from many queues in series, Ann. Appl. Prob. 1 (1991) 546–572.
[16] J. M. Harrison, Brownian models of queueing networks with heterogenous customer populations, in: *Stochastic Differential Systems, Stochastic Control Theory and Applications*, eds. W. Fleming and P.L. Lions (Springer, New York, 1988) pp. 147–186.
[17] J.M. Harrison and M.I. Reiman, Reflected Brownian motion on an orthant, Ann. Prob. 9 (1981) 302–308.
[18] J.M. Harrison and R.J. Williams, Multidimensional reflected Brownian motions having exponential stationary distributions, Ann. Prob. 15 (1987) 115–137.
[19] J.M. Harrison and R.J. Williams, Brownian models of open queueing networks with homogeneous customer populations, Stochastics 22 (1987) 77–115.
[20] L. Horvath, Strong approximation of open queueing networks, Math. Oper. Res. 17 (1992) 487–508.

[21] D.P. Johnson, Diffusion approximations for optimal filtering of jump processes and for queueing networks, Ph.D. Dissertation, Department of Mathematics, The University of Wisconsin, Madison (1983).
[22] O. Kella and W. Whitt, Diffusion approximations for queues with server vacations, Adv. Appl. Prob. 22 (1990) 706–729.
[23] A. Mandelbaum, The dynamic complementarity problem, Graduate School of Business, Stanford University (1989).
[24] J.L. Pomarede, A unified approach via graphs to Skorohod's topologies on the function space *D*, Ph.D. Dissertation, Department of Statistics, Yale University (1976).
[25] M.I. Reiman, Open queueing networks in heavy traffic, Math. Oper. Res. 9 (1984) 441–458.
[26] A.V. Skorohod, Limit theorems for stochastic processes, Theor. Prob. Appl. 1 (1956) 261–290.
[27] W. Whitt, Weak convergence theorems for queues in heavy traffic, Ph.D. Dissertation, Cornell University (1968).
[28] W. Whitt, Preservation of rates of convergence under mappings, Z. Wahrscheinlichkeitstheorie verw. Gebiete 29 (1974) 39–44.
[29] W. Whitt, The continuity of queues, Adv. Appl. Prob. 6 (1974) 175–183.
[30] W. Whitt, Some useful functions for functional limit theorems, Math. Oper. Res. 5 (1980) 67–85.