The Annals of Probability
1991, Vol. 19, No. 4, 1463–1519

STOCHASTIC DISCRETE FLOW NETWORKS: DIFFUSION APPROXIMATIONS AND BOTTLENECKS

BY HONG CHEN[1] AND AVI MANDELBAUM[2]

University of British Columbia and New Jersey Institute of Technology, and Stanford University and Technion–Israel Institute of Technology

Diffusion approximations for stochastic congested networks, both open and closed, are described in terms of the networks' bottlenecks. The approximations arise as limits of functional central limit theorems. The limits are driven by reflected Brownian motions on the nonnegative orthant (for open networks) and on the simplex (for closed ones). The results provide, in particular, invariance principles for Jackson's open queueing networks, Gordon and Newell's closed networks and some of Spitzer's finite particle systems with zero-range interaction.

Contents

1. Introduction. The present paper is concerned with diffusion approximations for congested discrete-flow networks in a stochastic environment. The network is represented by a graph and the discrete-flow is the motion of indistinguishable particles that individually traverse arcs and locally interact at nodes. One distinguishes between closed networks, in which the number of circulating particles is constant in time, and open networks, in which the number fluctuates due to the flow of particles in and out of the system. Congestion in a closed network increases with the number of circulating

Received March 1989; revised April 1990.

[1] Research supported in part by NSF Grant ECS-86-58157 and by a grant from AT & T Bell Laboratories.

[2] Research supported in part by the Alon Fellowship, the Israeli Center for Absorption in Science, the Bat-Sheva Foundation, the Technion V.P.R. Fund–S. & M. Wein Research Fund and the Semiconductor Research Corporation Grant 84-01-046.

AMS 1980 *subject classifications*. Primary 60F17, 60K25, 60J70, 90B10; secondary 60F15, 60K30, 60K35, 60G99, 90B15, 90B22, 90B30, 90C35, 93B99.

Key words and phrases. Flow networks, bottlenecks, fluid approximations, diffusion approximations, sample path analysis, queueing networks, heavy traffic, oblique reflection, reflected Brownian motions on the orthant and on the simplex.

particles, while congestion in an open network increases with the rate at which particles arrive. Our main results, stated formally as functional central limit theorems (FCLT's), provide diffusion approximations for congested discrete-flow networks, both open and closed. Our findings quantify the common belief that bottlenecks, loosely described as the heavily congested nodes, dominate the performance of the network.

1.1. For an historical account and further references on the diffusion approximations considered here, readers are referred to Whitt (1974), Lemoine (1978), Flores (1985) and Harrison (1988). They all discuss diffusion approximations in the context of queueing networks, which we also do starting at Section 2. The present work directly extends, streamlines or supports the studies by Iglehart and Whitt (1970), Johnson (1983), Reiman (1984), Kelly (1984), Goodman and Massey (1984), Harrison and Williams (1987) and Harrison, Williams and Chen (1990).

A salient feature of diffusion approximations is that they require no information about the model's stochastic primitives beyond the first and second moments of their distributions. This consequence of the invariance principle enables one to approximate non-Markovian nonparametric extensions of the Markovian exponential models considered, among others, by Jackson (1963) (who modelled job shop systems in manufacturing), Gordon and Newell (1967) (probably motivated by transportation systems in engineering), Whittle (1967, 1968) (whose terminology is suggestive of population models in biology), Spitzer (1970) (Section 2a, inspired by statistical mechanics models in physics) and Moore (1971) (whose domain of application is large-scale time-sharing systems in computer science). These references, all of which constitute pioneering treatments or seminal contributions in their corresponding areas, amply demonstrate the scope and potential applicability of the results reported in this paper. Thus, *nodes* and *particles* could have been replaced respectively by machines and production units, highways and vehicles, colonies and members, sites and particles, terminals and computer programs, economies and commodities, service stations and customers and more.

1.2. The stochastic processes that arise here as diffusion approximations are closely related to, or are themselves, multidimensional diffusions of a type known as reflected Brownian motions (RBM's). Our representation of RBM's originates in the work of Skorohod (1961) and is due to Harrison and Reiman (1981). For closed networks, RBM's approximate the temporal behaviour of the number of particles present at each node as a fraction of the total, hence their state space is the nonnegative unit simplex. For open networks, RBM's approximate absolute numbers, hence their state space is the nonnegative orthant. In both cases the state space is a polyhedron. The RBM evolves in the relative interior of this polyhedron as a Brownian motion, and from each of the facets of the polyhedron it is reflected instantaneously in a direction pointing into its relative interior.

In our context, RBM's arise only as diffusion approximations. However, RBM's have intrinsic interest both as challenging mathematical objects and as building blocks for *continuous-flow* networks. For example, Harrison (1985), Chapter 2, represents some simple continuous flow models in terms of RBM's. Varadhan and Williams (1985) and Williams (1987) address existence and uniqueness of RBM's within the framework of Stroock and Varadhan (1979); the first paper analyzes a two-dimensional general driftless RBM, while the second allows arbitrary dimension but imposes a certain skew-symmetry condition on the data. Harrison and Reiman (1981a) and Chen and Mandelbaum (1991b) are concerned with the sample-path construction of RBM's; the latter also relates them to continuous-time Leontief-substitution systems in economics. Harrison and Williams (1987) and Harrison, Williams and Chen (1990) analyze the stationary distribution of RBM's (see Subsection 8.7) and propose a procedure of actually fitting RBM's to open and closed networks, respectively. A last example is Reiman and Williams (1988) who focus on the behaviour of RBM's on the boundary of their state space.

1.3. The network model will be discussed within a framework that distinguishes three aggregation levels of *time* and *state space*. At a microscopic (lowest) level, full disclosure of details is required in order to represent the flow of individual particles. This is the level at which we set up our model of a stochastic discrete flow network in Section 2. In contrast, minimum data suffices at the macroscopic (highest) level, where the network is adequately approximated by a deterministic fluid model that captures its long-run average behaviour. Fluid models are described in Section 3, following Chen and Mandelbaum (1991a), hereafter abbreviated to CMa. Because the microscopic and macroscopic levels are far apart, it is natural to interpolate a mesoscopic (intermediate) level between them. To this end, deviations of the microscopic model from its fluid approximation are quantified by stochastic diffusion approximations for congested closed networks in Section 4, and for arbitrary open networks in Section 6. All our diffusion approximations are driven by RBM's which, for reasons that will become clear, are conveniently introduced as early as Section 3. The results stated in Section 4 are proved in Section 5. In Section 7 we extend the scope of Sections 4 and 6 to cover networks with several *prioritized* types of particles circulating in them. We conclude in Section 8 with some commentary, extensions and directions for future research.

The complexity of stochastic networks renders any comprehensive treatise of them somewhat cumbersome. An attempt has been made, therefore, to facilitate both the exposition and the access to main ideas (perhaps at the cost of not pursuing the most efficient presentation.) We hope to have facilitated the exposition by basing most arguments on Skorohod's (1956) representation theorem, thus reducing proofs of weak convergence for stochastic processes to the analysis of individual sample paths, as in Johnson (1983). In regard to main ideas, they are exposited in Sections 1–4 and are accessible without the prerequisite to delve into the proofs given later.

1.4. Formally described, diffusion approximations arise from a rescaling procedure that accelerates the time and aggregates the states of the microscopic model. The procedure amounts to a FCLT (Theorems 4.1, 6.1, 7.1), which accounts for the stochastic Gaussian nature of its outcome. A different rescaling procedure, that amounts to a functional strong law of large numbers (FSLLN), leads in CMa to the deterministic fluid approximations described in Subsections 3.1–3.2.

Speaking less formally, one uses microscopic, diffusion or fluid scales in order to highlight the dominating phenomena at each of the three levels of aggregation. This hierarchy of levels at which mathematical modelling is exercised has been fundamental, of course, across scientific disciplines [cf. Woods (1975)]. It has also contributed to the understanding and the design of managerial decision support systems. For example, a widely used hierarchy due to Anthony (1965) suggests that models at the microscopic, mesoscopic and macroscopic levels ought to support decision making over the operational short run, the tactical medium run and the strategic long run, respectively.

1.5. Let us describe an example which substantiates the discussion in Subsections 1.3–1.4 in the context of closed networks. [The example is a version of what Spitzer (1970) calls an interacting particle system with zero-range interaction, but we impose no assumptions of exponentiality.] Consider n identical particles circulating among J nodes of a graph. The nodes are indexed by $j = 1, \dots, J$, and each is equipped with a clock. Clock j at node j freezes when the node is empty and it rings at i.i.d. intervals with unit mean when one or more particles occupy the node. Suppose that the initial configuration of the n particles is drawn from a uniform distribution over all possible configurations. Independently thereafter, at each ring of clock j, a particle leaves node j and is placed immediately at node k with probability p_{jk}. The description thus far has been at the microscopic level. To illustrate the macroscopic and diffusion levels, let us assume for simplicity that the matrix $P = [p_{jk}]$ is irreducible *doubly* stochastic. We shall single out the J-dimensional stochastic process $Q^n = \{Q^n(t),\ t \geq 0\}$ whose jth coordinate at time t, $Q_j^n(t)$, represents the number of particles present at node j at time t.

The fluid approximation to the example, formally introduced in Section 3, is described in terms of J buffers. The buffers are connected by pipes to form a network within which fluid circulates. The transition of fluid between buffers is assumed to be instantaneous. There is a total of one unit of fluid and the initial configuration of fluid in the buffers is uniformly distributed on the J-dimensional unit simplex. Deterministically thereafter, each buffer releases fluid at a constant unit rate and a fraction p_{jk} of the fluid released from buffer j flows into the pipe leading directly to buffer k. Such dynamics turn out to maintain a constant fluid level at all the buffers at all times. This constant fluid configuration is the almost sure limit of $Q^n(nt)/n$ as $n \to \infty$. Generally, one arrives at the fluid approximation by increasing the number of particles n to infinity, while simultaneously accelerating time and aggregating space both by a factor of n (See the "bar" representation in Subsection 4.7.)

The diffusion approximation, described in Section 4, is obtained via rescaling time by a factor of n^2 and space by a factor of n (see the "hat" representation in Subsection 4.7.) In particular, when the J clocks are independent and the i.i.d. times between their successive rings have unit variance, as $n \to \infty$ the process $Q^n(n^2t)/n$ converges weakly to an RBM supported on the J-dimensional unit simplex. This RBM is stationary with initial state that is uniformly distributed on the simplex. It evolves inside the simplex like some J-dimensional Brownian motion. Upon reaching the facet of the simplex where the jth coordinate vanishes, $j = 1, \dots, J$, the RBM is reflected instantaneously towards the interior of the simplex with a direction of reflection that equals the jth row of the matrix $I - P$.

1.6. We now group, for convenience, the notations and conventions used throughout the paper.

VECTORS AND MATRICES. Vectors are understood to be column vectors. The transpose of a vector or a matrix is obtained by adding to it a prime. The J-dimensional Euclidean space is denoted by R^J, its nonnegative orthant by R^J_+ and the unit vectors in R^J by e^j, $j = 1, \dots, J$. When the letter e represents a vector, it is the vector of ones.

Let $x = (x_1, \dots, x_J)' \in R^J$ and $\alpha \subseteq \{1, \dots, J\}$. Then the scalar $|\alpha|$ denotes the cardinality of α and the vector $x_\alpha \in R^{|\alpha|}$ is the restriction of x to its coordinates with indices in α. Similarly, the matrix $P_{\alpha\beta}$ is the submatrix of a matrix P, obtained by choosing the elements with row-indices in α and column-indices in β; $P_{\alpha\alpha}$ will be abbreviated to P_α. The vector $|x| \in R^J$ stands for $|x| = (|x_1|, \dots, |x_J|)'$, while $\|x\| = \max_{1 \le j \le J}|x_j|$. Vector (in)equalities are to be interpreted componentwise. In particular, $x > y$ means that each coordinate of x is strictly greater than the corresponding y coordinate.

The spectral radius of a square matrix A is denoted by $\sigma(A)$. For a vector $\mu = (\mu_1, \dots, \mu_J)'$, the matrices diag($\mu$) and diag($\mu^{-1}$) are the $J \times J$ diagonal matrices with diagonal elements $(\mu_1, \dots, \mu_J)$ and $(1/\mu_1, \dots, 1/\mu_J)$, respectively.

SCALAR FUNCTIONS. The positive part of a scalar r is denoted by $r^+ = \max\{0, r\}$ and the integer part, namely the largest integer not exceeding r, by $\lfloor r \rfloor$. The indicator function of a set S is the function $1[S]$ that equals 1 when S prevails and 0 otherwise. The delta function δ_{jk} stands for $\delta_{jk} = 1$ when $j = k$ and $\delta_{jk} = 0$ otherwise. A real-valued nondecreasing function $y(t)$ has a point of increase (or simply increases) at $t \ge 0$ if $y(t + \varepsilon) > y(t-)$ for all $\varepsilon > 0$, with the convention $y(0-) = y(0)$.

VECTOR FUNCTIONS. An operation of a scalar function on a vector is to be interpreted coordinatewise. A vector-valued function is nondecreasing (nonincreasing) when all its components are nondecreasing (nonincreasing). The *composition* $\{x(y(t)),\ t \ge 0\}$ of x: $[0, \infty) \to R^J$ with y: $[0, \infty) \to R^J_+$ is the function from $[0, \infty)$ to R^J whose jth coordinate is the real-valued function

$\{x_j(y_j(t)),\ t \geq 0\}$, $j = 1, \dots, J$. In particular, the jth coordinate of $\{x(\mu t),\ t \geq 0\}$, $\mu = (\mu_1, \dots, \mu_J)' \in R_+^J$ is $\{x_j(\mu_j t),\ t \geq 0\}$.

FUNCTION SPACES. Denote by D^J the space of J-dimensional RCLL (right-continuous with left limits) functions, namely the R^J-valued functions on $[0, \infty)$ that are right-continuous on $[0, \infty)$ with finite left limits on $(0, \infty)$. We distinguish in D^J the following subsets: $D_0^J = \{x \in D^J\colon x(0) \geq 0\}$; $D_>^J = \{x \in D_0^J\colon e'x(t) > 0 \text{ for all } t \geq 0\}$; $C^J, C_0^J, C_>^J$ stand for the subset of continuous functions in the respective $D_\cdot^J$-sets; B^J, B_0^J, $B_>^J$ consist of the corresponding $D_\cdot^J$-functions which have bounded variation over finite intervals (locally of bounded variation). Finally, $D_\cdot^J[0, T)$ denotes the restrictions to $[0, T)$ of functions in $D_\cdot^J$; $C_\cdot^J[0, T)$ and $B_\cdot^J[0, T)$ are to be interpreted similarly.

For any $x \in D^J$ and $t > s > 0$, the uniform norm of x on the interval $[s, t]$ is the (necessarily finite) quantity

$$\|x\|_{[s,t]} = \sup_{s \leq u \leq t} \|x(u)\|,$$

with $\|x\|_{[0,t]}$ abbreviated to $\|x\|_t$.

CONVERGENCE. A sequence of functions in D^J converges u.o.c. if it converges uniformly on compact subsets of the semiclosed interval $[0, \infty)$. The sequence converges u.o.c. in $t > 0$ if the convergence is uniform on compact subsets of the open interval $(0, \infty)$.

Convergence in probability is denoted by $\rightarrow_p$. Convergence in distribution and weak convergence are used interchangeably and both are denoted by $\rightarrow_d$. Let $\{X_n\}$ be a sequence of stochastic processes with sample paths in D^J. Billingsley (1968) is the standard reference for results on the weak convergence of $\{X_n\}$, when $J = 1$ and time is restricted to $[0, 1]$. Extensions to $J \geq 1$ and to the time set $[0, \infty)$ have also become standard [see, e.g., Whitt (1980) or Chapter 3 in Ethier and Kurtz (1986)]. We further use weak convergence on $t > 0$ to denote weak convergence in $D^J(0, \infty)$, ($t = 0$ excluded). Whitt (1980) provides the modifications required to account for the exclusion of the origin [Resnick (1987) is an alternative textbook source].

A final comment is that, in this paper, weak convergence and convergence in probability are exclusively in D^J. However, all the stochastic processes that arise as weak limits have continuous sample paths, at least on $(0, \infty)$. Consequently, u.o.c. convergence of RCLL functions suffices for our purposes and there is no need to either introduce, discuss or use convergence with respect to any of Skorohod's topologies [see, e.g., Pollard (1984) who considers $[0, \infty)$]. Similarly, all the limits in probability that arise here are continuous on $[0, \infty)$. In this case, $X_n \rightarrow_p X$ with X continuous is equivalent to $\|X_n - X\|_t \rightarrow_p 0$ for all $t \geq 0$.

2. The network model. From now on we find it convenient to switch to the terminology of queueing networks and talk in terms of *service stations* and *customers* rather than *nodes* and *particles*, respectively. (This is, of course, only one out of many options, as examplified by the wealth of applications outlined in Subsection 1.1.) Our network, hereafter, consists of J interconnected service stations indexed by $j = 1, \dots, J$. Each station j constitutes a server, called server j, who is dedicated to serve customers present at station j on a first-come–first-served basis. Customers are indistinguishable and they arrive at a station either exogenously or from other stations. Upon completion of service at a station, a customer either leaves the network or transfers to another station in anticipation of additional service.

2.1. We take as primitive a probability space (Ω, F, P) on which the following random elements are defined: For $j = 1, \dots, J$, $Q_j(0)$ is a nonnegative integer-valued random variable, $u_j = \{u_j(n),\ n \geq 1\}$ a sequence of i.i.d. nonnegative random variables with unit mean and variance a_j^2, $v_j = \{v_j(n),\ n \geq 1\}$ a sequence of i.i.d. nonnegative random variables with unit mean and variance b_j^2 and $r^j = \{r^j(n),\ n \geq 1\}$ a sequence of i.i.d. random vectors with distribution supported on the set $\{0, e^1, \dots, e^J\}$ and specified by $\Pr\{r^j(n) = e^k\} = p_{jk}$, $k = 1, \dots, J$, and $\Pr\{r^j(n) = 0\} = 1 - \sum_{k=1}^{J} p_{jk}$. The random elements $Q(0) = (Q_1(0), \dots, Q_J(0))'$, u_j, v_j and r^j, $j = 1, \dots, J$, are assumed to be mutually independent.

The jth component $Q_j(0)$ of $Q(0)$ models the *initial queue length* at station j. The exogenous interarrival times to station j are constructed from the sequence u_j as follows: a nonnegative vector $\lambda^0 = (\lambda_1^0, \dots, \lambda_J^0)'$ is assumed given and the time between the $(n-1)$th and nth exogenous arrival to station j is taken to be $u_j(n)/\lambda_j^0$. Thus, the exogenous interarrival times to station j are i.i.d. with mean $1/\lambda_j^0$ and coefficient of variation (standard deviation divided by mean) that equals a_j. (When $\lambda_j^0 = 0$, the convention is that there are no exogenous arrivals to station j.) The service times are constructed similarly from the v_j's and a given positive vector $\mu = (\mu_1, \dots, \mu_J)'$. Specifically, the duration of the nth service performed by server j is $v_j(n)/\mu_j$. Thus, the service durations at station j are i.i.d. with mean $1/\mu_j$ and coefficient of variation b_j. Finally, the random sequence r^j models the routing mechanism enforced at station j as follows: the customer who completes the nth service at station j joins immediately station k if $r^j(n) = e^k$, $k = 1, \dots, J$, and leaves the network if $r^j(n) = 0$.

2.2. For $j = 1, \dots, J$, introduce

2.2.A $$U_j(n) = u_j(1) + \cdots + u_j(n), \qquad n \geq 1; \qquad U_j(0) = 0;$$

$$A_j^0(t) = \max\{n \geq 0: U_j(n) \leq t\}; \qquad A_j^0 = \{A_j^0(t), t \geq 0\};$$

2.2.B $$V_j(n) = v_j(1) + \cdots + v_j(n), \qquad n \geq 1; \qquad V_j(0) = 0;$$

$$S_j^0(t) = \max\{n \geq 0: V_j(n) \leq t\}; \qquad S_j^0 = \{S_j^0(t), t \geq 0\};$$

2.2.C $$R^j(n) = r^j(1) + \cdots + r^j(n), \qquad n \geq 1; \qquad R^j = \{R^j(n), n \geq 1\}.$$

Then let

2.2.D $A^0 = (A_1^0, \dots, A_J^0)'$; $\quad S^0 = (S_1^0, \dots, S_J^0)'$; $\quad R = [R^1, \dots, R^J]$.

(A^0 and S^0 are J-dimensional vector-valued processes, the coordinates of which are independent renewal processes; R is a $J \times J$-dimensional matrix-valued sequence, the columns of which are independent multinomial sequences.) Now recall the convention of composition between vector-valued functions, introduced in Subsection 1.6, and define the *exogenous arrival* process by $A = \{A(t) = A^0(\lambda^0 t),\ t \geq 0\}$ and the *service* process by $S = \{S(t) = S^0(\mu t),\ t \geq 0\}$. The jth coordinate $A_j(t)$ of A at time t represents the total number of exogenous arrivals to station j during the time interval $(0, t]$. Similarly, $S_j(u)$ models the total number of services performed by server j during its first u units of busy time. The (j, k)th coordinate $R_j^k(n)$ of the nth element in the *routing* sequence R indicates the number of customers, among the first n served at station k, which are routed directly to station j.

2.3. An application of the strong law of large numbers to A_j, S_j and R^j reveals, respectively, that λ_j^0 represents the long-run exogenous arrival rate of customers to station j, μ_j is the service rate of server j and p_{jk} the long-run fraction of customers that switch directly to station k after being served by server j. In accordance with these interpretations, λ^0 will be called the *exogenous arrival-rate* vector, μ the *service-rate* vector and the matrix $P = [p_{jk}]$ the *switching matrix*. The triplet (λ^0, P, μ) determines long-run average performance. Indeed, the *inflow capacity* vector of the network is the maximum solution $\lambda = (\lambda_1, \dots, \lambda_J)'$ to the *traffic equations*

$$\lambda_j = \lambda_j^0 + \sum_{k=1}^{J} (\lambda_k \wedge \mu_k) P_{kj},$$

which reads in vector form

$$\lambda = \lambda^0 + P'(\lambda \wedge \mu). \tag{2.1}$$

The traffic equations (2.1) first appeared in Massey (1981) in the context of open networks. A maximum solution λ exists by Theorem 3.1 in CMa, where λ_j is interpreted as a least upper bound to the long-run *actual* arrival rate that can be realized at station j. Define the *traffic intensity* of station j to be

$$\rho_j = \lambda_j / \mu_j, \qquad j = 1, \dots, J,$$

and call station j a *bottleneck* station if $\rho_j \geq 1$, *strict bottleneck* if $\rho_j > 1$, *nonbottleneck* if $\rho_j < 1$ and a *balanced* station if $\rho_j = 1$. Hereafter,

2.3.A The set of nonbottleneck, balanced and strict bottleneck stations will be denoted by α, β and γ, respectively.

Bottlenecks, traffic intensities and other related concepts are discussed in CMa, especially Section 4 which is devoted to an exposition of network properties that are solely determined by the triplet (λ^0, P, μ). There are several important such properties, and when discussing them it will be convenient to identify the network with its corresponding triplet. In particular, a network (λ^0, P, μ) is called *open* if the spectral radius $\sigma(P)$ of the switching matrix P is strictly less than unity. A network is *closed* if P is stochastic and $\lambda^0 = 0$ and *irreducible closed* if, furthermore, P is an irreducible matrix. The population of customers in an open network fluctuates in time and each customer eventually leaves after having visited a finite number of stations on its route. (In fact, the average number of stations visited by a customer is finite.) In contrast, the population that circulates within a closed network is unchanging and each customer eventually visits all stations when the network is irreducible. Another distinction between open and closed networks is that in a closed network $\lambda \leq \mu$ must prevail. Consequently, there are no strict bottlenecks in a closed network and all the bottlenecks are balanced.

2.4. We are about to construct the *busy time* process $B = \{B(t),\ t \geq 0\}$, which is a J-dimensional stochastic process whose jth coordinate at time t, $B_j(t)$, measures the cumulative amount of time that server j has been busy during the time interval $(0, t]$. In terms of B, the total number of arrivals to station j up to time t (either exogenously or from the other stations) equals $A_j(t) + \sum_{k=1}^{J} R_j^k\{S_k[B_k(t)]\}$, $t \geq 0$. Consequently, the *queue-length* process $Q = \{Q(t) = (Q_1(t), \dots, Q_J(t))',\ t \geq 0\}$, whose jth coordinate $Q_j(t)$ represents the number of customers queued in station j at time t, must be

$$(2.2)\quad Q_j(t) = Q_j(0) + A_j(t) + \sum_{k=1}^{J} R_j^k\{S_k[B_k(t)]\} - S_j[B_j(t)], \qquad t \geq 0.$$

To completely specify the dynamics of the network, we assume that servers work at full rate whenever their queues are not empty. [In the queueing literature this is referred to as a work-conserving or non-idling service discipline; see, e.g., Wolff (1970).] The non-idling constraint on server j takes the form

$$(2.3)\qquad B_j(t) = \int_0^t 1[Q_j(u) > 0]\, du, \qquad t \geq 0.$$

The existence of the queue-length process Q and the busy-time process B is established in Section 7 of CMa, where it is verified that there exists a unique nonnegative pair (Q, B) that satisfies simultaneously (2.2) and (2.3) for all $j = 1, \dots, J$ and $t \geq 0$.

It turns out that the fluid and diffusion approximations of the queue length process are supported on the set of bottleneck stations. This is demonstrated for fluids in CMa, for diffusions associated with closed models in (4.2) and with open models in (6.2). The phenomenon, first established for acyclic networks in Theorem 1c of Iglehart and Whitt (1970), quantifies the negligible effect of nonbottlenecks, relative to bottlenecks, on queue buildups and it might provide a potentially useful reduction in the complexity of the model. Our results for open networks supplement Reiman (1984) and Johnson (1983). Both employ the representation (2.2)–(2.3) due to Harrison.

2.5. Important performance measures, in addition to queue lengths and busy times, are workloads and sojourn times which will now be defined. The *workload* process $W = (W_1, \dots, W_J)'$ is the J-dimensional stochastic process whose jth coordinate at time t, $W_j(t)$, models the amount of work, measured in units of service time, that awaits server j at his station at time t. Recalling that the duration of the nth service performed by server j is $v_j(n)/\mu_j$, one deduces from (2.2) and 2.2.B that

$$W_j(t) = V_j\big[Q_j(t) + S_j\big(B_j(t)\big)\big]/\mu_j - B_j(t), \qquad t \geq 0, \tag{2.4}$$

for $j = 1, \dots, J$. The representation (2.4) gives rise to a notable relation between the approximations (both fluid and diffusion) of the queue-length and the workload processes: as demonstrated by Theorem 7.1 in CMa for fluids and by (4.5) and (6.3) for diffusions, one process is a *linear* function of the other. We shall further comment on this phenomenon in the next subsection.

2.6. Let $h = (h_1, \dots, h_J)'$ have nonnegative integer-valued coordinates and fix a station j. Suppose the switching matrix P is such that it is possible for a customer, upon entering j and before returning to it, to follow a route of length $e'h$ in which station k is visited h_k times, $k = 1, \dots, J$. Then h will be called a *visit vector* that is accessible from j, or *j-accessible* for short; a customer who follows such a route is said to *follow h upon entering j*. (Note that $h_j = 1$ for j-accessible vectors h because the route starts at j and it terminates before returning to j.) The time it takes a specific customer to follow h is called his *sojourn time along h*. For a j-accessible h, let $D_{j,h}$ denote the stochastic process whose state at time t, $D_{j,h}(t)$, is the sojourn time along h of the first customer who follows h upon entering j after time t. We now proceed with the formal definition of $D_{j,h}$.

A route that is *associated* with a j-accessible visit vector $h = (h_1, \dots, h_J)'$ constitutes a sequence of indices $i_0, i_1, \dots, i_{e'h}$ such that $i_0 = j$, $i_n \neq j$ for $n = 1, \dots, e'h - 1$, $p_{i_{n-1}, i_n} > 0$ for $n = 1, \dots, e'h$, and the cardinality of the set $\{n\colon i_n = k,\ n = 0, 1, \dots, e'h - 1\}$ equals h_k, $k = 1, \dots, J$. Consider the first customer who follows h upon entering j after time t. Denote by $\tau_{j,h,k,i}(t)$ the time this customer starts the ith visit to station k while following the route associated with h. A consequence of the first-come–first-served service discipline is that for a j-accessible h we have

$$D_{j,h}(t) = \sum_{k=1}^{J} \sum_{i=1}^{h_k} W_k\big[\tau_{j,h,k,i}(t)\big], \tag{2.5}$$

where the W_k's are the workloads defined in (2.4). For closed networks, we shall analyze the *sojourn time* process $D = \{D_{j,h}\}$, whose components constitute the set of $D_{j,h}$, where h is j-accessible, $j = 1, \dots, J$. The sojourn time process for open networks is modified in two ways. First, the h's are restricted to satisfy $h_i = 0$ for $i \in \gamma$ ($h_\gamma = 0$ for short), which means that their associated routes do not encounter strict bottlenecks. Second, one allows 0-accessible h's, where "station" 0 stands for the world external to the network and arrivals to station 0 are arrivals to the network. (The obvious interpretation of the condition $p_{0,i} > 0$, required when a route is associated with a 0-accessible h, is that $\lambda_i^0 > 0$.)

The diffusion limits of sojourn times are formulated in (6.13) for open networks [overlapping Reiman (1982, 1984), who attributes a key observation to Foschini] and in (4.9) for closed networks [as anticipated in Section 3 of Kelly (1984)]. From their derivation (see Subsection 5.6), the following remarkable facts emerge. Consider, for example, a route in a closed network which does not return to the station it started from. Such a route is associated with some j-accessible h with $h_j = 1$. It turns out that a diffusion approximation for the time to traverse this route is provided by the diffusion limit of $D_{j,h}$ which, in turn, equals the scalar product of h with the diffusion limit of the workload process. It is as if, in the diffusion scale, queues and workloads seen by a customer upon arriving at j freeze while the customer follows h. With nonbottleneck queues being empty in the diffusion limit, it follows that there is no delay while passing through nonbottlenecks [see (4.9)]. Also the *order* in which stations are visited along a route does not affect the diffusion approximation of the route's travel time. (For open networks this prevails as long as the route does not encounter strict bottlenecks, but it fails otherwise.) Furthermore, the diffusion limit of $D_{j,h}$, $h_\gamma = 0$, is independent of j in the sense that the limits of $D_{j,h}$ and $D_{k,h}$ coincide when h is both j-accessible and k-accessible, $j \neq k$. Finally, the dependence between the diffusion limits of sojourn times and queue lengths at stations is also linear, which is a manifestation of Little's law in the present setup [see Glynn and Whitt (1986)].

A comment on our use of the term sojourn time is in order. In the queueing literature, the sojourn time of a customer in an open network stands for the elapsed time between its arrival and eventual departure epochs. Natural standard analogs for closed networks are first passage times between or return times to stations. As observed by Reiman (1984), however, these standard concepts lead to stochastic processes whose limiting behaviour differs from that encountered here. For example, consider the first customer that arrives after time t to station j of an irreducible closed network. Define D_j to be the stochastic process whose state at t is the *return-time* to j of that customer (namely, the elapsed time between instants of successive arrivals to j). Then the standard rescaling of D_j (as in Section 4.2) typically fails to converge weakly to a continuous-path process, because customers who arrive close in time may have rather different routes, hence substantially different return times. In contrast, rescaled sojourn times along a specific route do converge to a continuous limit. (Generally, our results still apply to standard sojourn times and their analogs if and only if the entries of the switching matrix P are 0 or 1's; when P is irreducible, such a network must consist of stations in series if it is open and stations in a cycle if closed.)

2.7. A major intermediate step in the derivation of our results is the centering of the network's components around what might be called their asymptotic mean functions. Centering highlights the mathematical link among the three levels of aggregation discussed in Subsection 1.3. This link constitutes the oblique reflection mapping, to be introduced in Subsection 3.2 and a representation alternative to (2.2)–(2.3), which will now be presented. Define

$$\theta = \lambda^0 + P'\mu - \mu \tag{2.6}$$

and let

$$\begin{aligned}
\xi_j(t) &= \left[A_j(t) - \lambda_j^0 t\right] + \sum_{k=1}^{J} \left\{R_j^k[S_k(B_k(t))] - p_{kj}S_k(B_k(t))\right\} \\
&\quad + \sum_{k=1}^{J} p_{kj}\left[S_k(B_k(t)) - \mu_k B_k(t)\right] - \left[S_j(B_j(t)) - \mu_j B_j(t)\right], \\
X_j(t) &= Q_j(0) + \theta_j t + \xi_j(t), \\
Y_j(t) &= \mu_j\left[t - B_j(t)\right],
\end{aligned} \tag{2.7}$$

for $j = 1, \ldots, J$ and $t \geq 0$. Then the pair (Q, B) uniquely satisfies (2.2) and (2.3) if and only if the pair (Q, Y) uniquely satisfies the three relations

2.7.A $Q = X + [I - P']Y \geq 0$,
2.7.B Y is nondecreasing with $Y(0) = 0$ and
2.7.C Y_j increases only at times t when $Q_j(t) = 0$, $j = 1, \ldots, J$.

The representation 2.7.A–2.7.C plays a key role in our analysis. Its importance stems from the fact that both the fluid models [see the "bar" representation 4.7.A] and the diffusion models [see the "hat" representation 4.7.B] are its direct descendants.

3. Fluid models, reflection mappings and RBM's.

As already mentioned in Subsection 2.7, the oblique reflection mapping mathematically links the microscopic specification of the network in Section 2 with its macroscopic and diffusion approximations. The mapping was introduced for open models in Harrison and Reiman (1981a) and Reiman (1984). Chen and Mandelbaum (1991b) (CMb) extend its scope to cover closed models. Readers are referred to these works if they need proofs of the facts that will be outlined later. At the microscopic level, the mapping is implicit in 2.7.A–2.7.C. We shall now animate it by a linear fluid model which is, in fact, the macroscopic approximation alluded to before. Then the mapping will be used to construct diffusion approximations, namely open and closed reflected Brownian motions (RBM's). Open RBM's, which approximate open networks, are the subject of Harrison and Reiman (1981a) and Harrison and Williams (1987). Closed RBM's approximate closed networks and are treated in Harrison, Williams and Chen (1990) and CMb.

3.1. Consider a collection of J buffers indexed by $j = 1, \ldots, J$. The buffers have infinite storage capacity and they are interconnected by frictionless pipes to form a network within which a homogeneous fluid is circulating. Initially at time $t = 0$, the fluid level at buffer j is $Z_j(0) \geq 0$, $j = 1, \ldots, J$. Then fluid flows to each buffer either exogenously from the outside world or internally from other buffers. Exogenous inflow to buffer j is at a constant rate $\lambda_j^0 \geq 0$ and the release capacity of buffer j is given by $\mu_j > 0$. We postulate that fluid is released from all the buffers at the maximum rate possible. Thus, the release rate at buffer j is μ_j when the buffer is not empty and it coincides with the smaller of μ_j and the actual (exogenous plus internal) inflow rate when buffer j is empty. (Note that this description is not complete in that the internal, hence the actual, inflow rates are not known in advance. The mathematical formulation that will be given momentarily circumvents this difficulty). Finally, a fraction $p_{jk} \geq 0$ of the total outflow from buffer j is routed directly to buffer k, $k = 1, \dots, J$, and a fraction $1 - \sum_{k=1}^{K} p_{jk}$ leaves the network. To summarize, a nonnegative J-dimensional vector $Z(0)$ and a triplet (λ^0, P, μ), as in Subsection 2.3, constitute all the data required to specify the fluid network. It is natural to call such a network *open* when $\sigma(P) < 1$ and *irreducible closed* when both P is irreducible stochastic and $\lambda^0 = 0$.

3.2. If buffer j is never empty during the time period $[0, t]$, then its cumulative outflow up to time t amounts to $\mu_j t$. If the buffer does get empty, then some of its potential outflow is lost. Define $Y_j(t)$ to be the cumulative outflow from buffer j that may get lost due to its emptiness during $[0, t]$. Then the actual cumulative outflow from j up to t equals $\mu_j t - Y_j(t)$ and the fluid level in buffer $j = 1, \dots, J$ at time $t \geq 0$ must be

$$(3.1) \quad Z_j(t) = Z_j(0) + \lambda_j^0 t + \sum_{k=1}^{J} [\mu_k t - Y_k(t)] p_{kj} - [\mu_j t - Y_j(t)].$$

In vector notation, (3.1) reads

$$3.2.\text{A} \qquad Z(t) = X(t) + [I - P'] Y(t), \qquad t \geq 0.$$

Here, the *fluid level* process $Z = \{Z(t),\ t \geq 0\}$ is J-dimensional with coordinates $Z_j(t)$ as in (3.1), the *loss* process $Y = \{Y(t),\ t \geq 0\}$ has coordinates $Y_j(t)$ and the *linear netflow* process $X = \{X(t),\ t \geq 0\}$ is defined to be

$$(3.2) \qquad X(t) = Z(0) + (\lambda^0 + P'\mu)t - \mu t = Z(0) + \theta t, \qquad t \geq 0,$$

with $\theta = \lambda^0 + [P' - I]\mu$ as in (2.6). [The terminology "netflow" has been chosen because X is the difference between the vector of cumulative *potential* inflows $(\lambda^0 + P'\mu)t$ and *potential* outflows μt; "linear" signifies the linear dependence of the netflow on time and distinguishes it from the more general X's that will arise later.] The dynamics of the network are mathematically expressed by the following three constraints:

3.2.B Z is nonnegative,
3.2.C Y is nondecreasing with $Y(0) = 0$ and
3.2.D Y_j increases only at time $t \geq 0$ when $Z_j(t) = 0$, $j = 1, \dots, J$.

The roles of 3.2.B and 3.2.C are obvious. The constraint 3.2.D forces potential outflow to be lost *only* due to emptiness of buffers. It follows from the results in CMb that this is, in fact, equivalent to the release mechanism postulated previously, namely maximizing the outflow rates at all buffers at all times without exceeding μ. One can further show that there exists a unique pair (Y, Z) satisfying 3.2.A–3.2.D, thus providing a model representation of the *linear* fluid network. Let us remark for future reference that (Y, Z) depends on the data of the network only through X and P.

It turns out (Theorem 5.2 in CMa; see also Subsection 4.6) that a linear fluid network reaches equilibrium in the following sense. There exists a finite time τ after which the nonbottlenecks are always *empty*; actual inflow rates to the buffers then coincide with the inflow capacities that arise from the traffic equations (2.1), with (λ^0, P, μ) being the fluid network's parameters; (up to time τ, flow rates change in a piecewise linear fashion). In Theorem 7.1 of CMa (which is a FSLLN), the linear fluid network emerges as the almost-sure limit of a rescaling procedure that successively accelerates the time and aggregates the states of the discrete flow network. This justifies its use as a fluid approximation. (The rescaling is formalized by the "bar" convention introduced in Subsection 4.7.)

3.3. Consider the mapping by means of which (Y, Z) is obtained from X through 3.2.A–3.2.D (P is assumed fixed). The linearity of the netflow plays no role here and the mapping is indeed extendable to nonlinear X's. For some such X's the mapping can be still animated by a *nonlinear* fluid network (as in the introduction to CMa). For other X's, however, mainly its mathematical significance is retained. When so extended, the mapping will be referred to as the *oblique reflection mapping*. We now proceed with its formal definition and then outline some of its properties that will be used later on.

Let P be a $J \times J$ nonnegative matrix. A *regulator* of an $x \in D_0^J$ is an element $y \in D_0^J$ such that

3.3.A $z = x + [I - P']y \geq 0$,
3.3.B y is nondecreasing with $y(0) = 0$ and
3.3.C y_j increases only at time $t \geq 0$ when $z_j(t) = 0$, $j = 1, \dots, J$.

The oblique reflection mapping Ψ_P is defined on $x \in D_0^J$ for which a regulator *y exists and is unique*. We summarize such state of affairs by writing $\Psi_P(x) = y$ and saying either that x is in the *domain* of Ψ_P or that Ψ_P is *well-defined* at x. Now recall the u.o.c. conventions in Subsection 1.6, and let x belong to the domain of Ψ_P. Then Ψ_P is *continuous* at x if for any sequence $\{x_n\}$ in its domain that converges u.o.c. to x, the sequence $\{\Psi_P(x_n)\}$ converges u.o.c. to $\Psi_P(x)$. Harrison and Reiman (1981a) showed that

3.3.D If $\sigma(P) < 1$, then Ψ_P is well-defined and continuous on D_0^J,
which is useful for open networks. An analog for closed models reads
3.3.E If P is irreducible stochastic, then Ψ_P is well-defined on $B_>^J \cup C_>^J$ and is continuous on $C_>^J$.

[3.3.E is Theorem 2.6 in CMb.]

3.3.F Suppose that $\sigma(P) < 1$ and $x \in D^J_0$, or that P is irreducible stochastic and $x \in B^J_{>} \cup C^J_{>}$. Then $y = \Psi_P(x)$ if and only if for all $j = 1, \dots, J$ and $t \geq 0$,

$$y_j(t) = \min\big\{y'_j(t) \colon y' \in D^J,\ x + [I - P']y' \geq 0,$$
$$y' \text{ is nondecreasing with } y'(0) = 0\big\}.$$

Alternatively, y is the least among all the elements in D^J which satisfy 3.3.A–3.3.B.

[3.3.F combines the Appendix in Reiman (1984), Proposition 2.4 and Theorem 2.5 in CMb.]

Properties 3.3.A–3.3.C and facts 3.3.E–3.3.F can all be restated for elements in $D^J[0, T)$, $T > 0$ arbitrary. Doing so will be a necessity when uniqueness of a regulator can be guaranteed only up to some $T < \infty$. [An example is (5.18) where P is irreducible stochastic and $x \in B^J_0$ is guaranteed to satisfy $e'x(t) > 0$ only in $t \in [0, T)$, for some $T > 0$.] We shall then write $\Psi_P(x) = y$ on $[0, T)$ and say that Ψ_P is well-defined at x up to T.

3.3.G Let P be irreducible stochastic. Consider a sequence $\{x_n\}$ that converges u.o.c. to $x \in C^J_{>}$. Suppose that $y_n = \Psi_P(x_n)$ on $[0, T_n)$, where $T_n \to \infty$ as $n \to \infty$. Then $\{y_n\}$ converges u.o.c. to $y = \Psi_P(x)$.

[3.3.G is Lemma 2.8 in CMb.]

3.4. Let $X = \{X(t),\ t \geq 0\}$ be a J-dimensional Brownian motion starting at $X(0) \geq 0$. We shall write $X = \mathrm{BM}(\delta, \Lambda)$ to indicate that X has a drift vector δ and covariance matrix Λ. When $\sigma(P) < 1$, one relies on 3.3.D to define the sample paths of $Z = \{Z(t),\ t \geq 0\}$, the *open* RBM, by

3.4.A $$Z = X + [I - P']Y, \quad \text{where } Y = \Psi_P(X).$$

When P is irreducible stochastic, one defines the *closed* RBM also as Z in 3.4.A. Here, however, the definition is based on 3.3.E supplemented by the requirement that

3.4.B $$e'X(0) = 1, \qquad e'\delta = 0 \quad \text{and} \quad e'\Lambda = 0.$$

The conditions in 3.4.B are necessary and sufficient for X to have essentially all sample paths in C^J_1. We shall write $Z = \mathrm{RBM}(\delta, \Lambda, P)$ to denote an RBM, either open or closed, which arises from a Brownian motion $X = \mathrm{BM}(\delta, \Lambda)$ via Ψ_P. The main distinction between open and closed RBM's is their state space: For an open RBM it is the J-dimensional nonnegative orthant when Λ is of rank J. For a closed RBM, Λ is of rank $J - 1$ at the most and the state space is the unit simplex in R^J when the rank is $J - 1$.

3.5. A geometric interpretation of an RBM is the following: Z behaves like X in the relative interior of its state space; when it hits the relative interior of the boundary face $\{z_j = 0\}$, Z is reflected instantaneously back into its state space; the direction of reflection is the jth column of $I - P'$ and reflection is accomplished by the minimal increase in the jth coordinate Y_j of Y that maintains the nonnegativity of Z intact. Reflection from an intersection of faces is carried out along a nonnegatively weighted sum of directions by increasing several coordinates of Y simultaneously. For example, the jth and kth columns of $I - P'$ are used upon hitting the relative interior of the face $\{z_j = 0\} \cap \{z_i = 0\}$. The above geometric interpretation is rather loose. A tighter description is provided in Section 2 of Mandelbaum (1990).

An RBM has continuous sample paths, it is adapted to the Brownian motion X which generates it and is a Markov process with stationary transition probabilities. These facts are, respectively, a consequence of the following properties of the oblique reflection mapping: Assume that either $\sigma(P) < 1$ and $x \in C_0^J$, or that P is irreducible stochastic and $x \in C_1^J$; let (x, y, z) satisfy 3.3.A–3.3.C. Then:

3.5.A Both y and z are continuous functions of x.

3.5.B For every $t > 0$, the restrictions of y and z to $[0, t]$ depend only on the restriction of x to $[0, t]$.

3.5.C Let $T \geq 0$. Define $\vec{x}(t) = z(T) + x(T + t) - x(T)$, $\vec{y}(t) = y(T + t) - y(T)$ and $\vec{z}(t) = z(T + t)$ for $t \geq 0$. Then $\vec{y} = \Psi_P(\vec{x})$ and $\vec{z} = \vec{x} + [I - P']\vec{y}$.

4. Diffusion approximations for irreducible closed networks. We are now ready to formulate the limit theorem for irreducible closed queueing networks. Closed RBM's emerge from the theorem as diffusion approximations when congestion is heavy. Congestion in a closed network increases with the number of customers circulating in it. Thus, we analyze a sequence of closed networks indexed by their population size n and obtain a closed RBM in the limit as n increases indefinitely. While the number of stations is held fixed, other network specifications, such as μ and P, can vary with n. To simplify the presentation, however, only variations in μ will be analyzed here and extensions are relegated to Section 8.

4.1. The sequence of irreducible closed networks is indexed by $n = 1, \dots$. The nth network has an initial queue length vector $Q^n(0)$ and a triplet $(0, P, \mu^n)$ so that $e'Q^n(0) = n$, the matrix P is irreducible stochastic and μ^n is a positive vector. As in Subsection 2.2, the service process S^n of the nth network is constructed from S^0 by $S^n(t) = S^0(\mu^n t)$, the routing sequence is R and, of course, there are no exogenous arrivals. Finally, the queue length process Q^n and the busy-time process B^n are uniquely determined by (2.2)–(2.3), $Q^n(0)$, S^n and R. To summarize, in the nth network the population size equals n, the mean service time at station j is $1/\mu_j^n$, the standard deviation of the service time distribution equals b_j/μ_j^n and its coefficient of variation is still b_j. The convention of appending a superscript n to performance measures associated with the nth network will be maintained throughout.

4.2. In our theorem we assume that for some vector $c, \mu > 0$ and a random vector $\hat{Q}(0) \geq 0$, all of which are J-dimensional, the following limits exist as $n \to \infty$:

4.2.A $$n(\mu^n - \mu) \to c, \quad \text{hence } \mu^n \to \mu;$$

4.2.B $$\frac{1}{n} Q^n(0) \to_d \hat{Q}(0).$$

In the formulation of the theorem we use a J-dimensional driftless Brownian motion

4.2.C $$\hat{\xi} = \mathrm{BM}(0, \hat{\Lambda})$$

which starts at $\hat{\xi}(0) = 0$. The covariance matrix $\hat{\Lambda} = [\hat{\Lambda}_{jk}]$ is given by

4.2.D $$\hat{\Lambda}_{jk} = \lambda_j \left(1 + b_j^2\right) \delta_{jk} - \lambda_j b_j^2 p_{jk} - \lambda_k b_k^2 p_{kj} - \sum_{l=1}^{J} \lambda_l p_{lj} p_{lk} \left[1 - b_l^2\right],$$

where $\lambda = (\lambda_1, \dots, \lambda_J)'$ is the inflow capacity vector of the closed network $(0, P, \mu)$, as defined via (2.1). The limit theorem is a functional central limit theorem jointly for the sequences

$$\hat{Q}^n(t) = \frac{1}{n} Q^n(n^2 t), \qquad \hat{W}^n(t) = \frac{1}{n} W^n(n^2 t),$$

$$\hat{D}^n(t) = \frac{1}{n} D^n(n^2 t), \qquad \hat{B}^n(t) = \frac{1}{n}\left[B^n(n^2 t) - \rho^n n^2 t\right].$$

Here ρ^n is the traffic intensity vector of the network $(0, P, \mu^n)$ as defined in Subsection 2.3; Q^n is the queue length and B^n the busy-time processes associated with the nth network, W^n is its workload process given in Subsection 2.5 and $D^n = \{D^n_{j,h}\}$ is its sojourn time process from Subsection 2.6. (Recall that D^n is indexed by all j-accessible h's, $j = 1, \dots, J$, along which no returns to j are allowed, or formally, $h_j = 1$.) The time scale of the nth network is accelerated by a factor of n^2 while states are aggregated by a factor of n. In particular, $\hat{Q}^n_j$ is an accelerated version of the queue length at station j as a *fraction* of the total n.

THEOREM 4.1. *Consider the sequence of closed networks in Subsection* 4.1. *Assume* 4.2.A–4.2.B *hold and let* $\hat{\xi}$ *be the Brownian motion* 4.2.C. *Then the weak convergence*

(4.1) $$\left(\hat{Q}^n, \hat{W}^n, \hat{B}^n, \hat{D}^n\right) \to_d \left(\hat{Q}, \hat{W}, \hat{B}, \hat{D}\right) \quad \textit{in } t > 0,$$

holds as $n \to \infty$. The limit is described for $t > 0$ by

$$\hat{Q}_\alpha = 0, \tag{4.2}$$

$$\hat{Q}_\beta = X + [I - \tilde{P}']Y, \tag{4.3}$$

$$X(t) = X(0) + \hat{\xi}_\beta(t) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{\xi}_\alpha(t) + [\tilde{P}' - I]c_\beta t, \tag{4.4}$$

$$X(0) = \hat{Q}_\beta(0) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{Q}_\alpha(0), \tag{4.5}$$

$$\tilde{P} = P_\beta + P_{\beta\alpha}[I - P_\alpha]^{-1}P_{\alpha\beta}, \tag{4.6}$$

$$Y = \Psi_{\tilde{P}}(X), \tag{4.7}$$

$$\hat{W} = \operatorname{diag}(\mu^{-1})\hat{Q}, \tag{4.8}$$

$$\hat{D}_{j,h} = \sum_{k \in \beta} \frac{h_k}{\mu_k}\hat{Q}_k = h\hat{W}, \tag{4.9}$$

$$\hat{B}_\alpha = \operatorname{diag}(\mu_\alpha^{-1})[I - P'_\alpha]^{-1}\big[\hat{Q}_\alpha(0) + \hat{\xi}_\alpha - P'_{\beta\alpha}Y + P'_{\beta\alpha}\, d_\beta t\big], \tag{4.10}$$

$$d_\beta = c_\beta - \left(\min_{i \in \beta} \frac{c_i}{\lambda_i}\right)\lambda_\beta,$$

$$\hat{B}_\beta = -\operatorname{diag}\big(\mu_\beta^{-1}\big)Y. \tag{4.11}$$

4.3. *Remarks.*

Remark 1. In addition to the parameters μ, c, P and $\hat{Q}(0)$, the Brownian motion $\hat{\xi}$ is the only data required to determine uniquely the left-hand sides of the equalities (4.2)–(4.11). The diffusion limits associated with the bottleneck subnetwork β are functions merely of X, which is identified in (4.4) as a $|\beta|$-dimensional Brownian motion starting at $X(0)$ with drift

$$\delta = [\tilde{P}' - I]c_\beta,$$

and covariance

$$\Lambda = \hat{\Lambda}_\beta + \hat{\Lambda}_{\beta\alpha}[I - P_\alpha]^{-1}P_{\alpha\beta} + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{\Lambda}_{\alpha\beta} + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{\Lambda}_\alpha[I - P_\alpha]^{-1}P_{\alpha\beta}.$$

Here, $X(0)$ is defined in (4.5), $\tilde{P}$ in (4.6), c_β in 4.2.A and $\hat{\Lambda}$ in 4.2.D.

Remark 2. Lemmas 4.4 and 4.3 in CMa guarantee, respectively, that $[I - P'_\alpha]$ is invertible and that $\tilde{P}$ in (4.5) is irreducible stochastic. One can check that 3.4.B also holds, thus identifying $\hat{Q}_\beta$ in (4.3) as the *closed* RBM

$$\hat{Q}_\beta = \mathrm{RBM}(\delta, \Lambda, \tilde{P}).$$

Remark 3. The random variable $\hat{Q}_\alpha(0)$ represents the limiting fraction of customers initially at the nonbottlenecks. If $\hat{Q}_\alpha(0) \neq 0$, then (4.1) does *not* hold for $t \geq 0$ because the processes involved are not right-continuous at $t = 0$. Indeed, substituting $t = 0$ in (4.3) then contradicts (4.5) ($Y(0) = 0$) and (4.10) with $t = 0$ disagrees with $\hat{B}^n(0) = 0$. [Weak convergence at $t = 0$ separately does hold with the following limits: $\hat{Q}(0)$ is given in 4.1.B, (4.8)–(4.9) determine $\hat{W}(0)$ and $\hat{D}(0)$, and $\hat{B}(0) = 0$.]

Remark 4. The weak convergence of $\hat{B}^n_\beta$ to $\hat{B}_\beta$ is actually in the functional space $C^{|\beta|}$. In fact, if $\hat{Q}_\alpha(0) = 0$, then (4.1)–(4.11) *all* hold for $t \geq 0$.

Remark 5. The set γ is lacking from the statement of the theorem because closed networks have no strict bottlenecks. The congestion measures of the set of nonbottlenecks α vanish in the limit. An important consequence is that the set of balanced stations β can be approximated, for large n, by the autonomous irreducible closed network $(0, \tilde{P}, \mu_\beta)$. [A more formal description is provided in Subsection 4.8.]

Remark 6. The covariance matrices $\hat{\Lambda}$ in 4.2.D and Λ in Remark 1 depend on the service time distributions only through the coefficients of variation b_j (rather than through means and variances separately).

Remark 7. In matrix form

$$\hat{\Lambda} = \Gamma[I + \Delta] - \Gamma\Delta P - P'\Delta\Gamma - P'[I - \Delta]\Gamma P,$$

where

$$\Gamma = \operatorname{diag}(\lambda) \quad \text{and} \quad \Delta = \operatorname{diag}\left(b_1^2, \dots, b_J^2\right).$$

In concert with the approximation suggested in Remark 5, Λ has that same representation with $\tilde{P}$ replacing P.

Remark 8. If service rates at the bottleneck stations do not vary with n ($\mu^n_\beta = \mu_\beta$), then X is driftless.

Remark 9. A consequence of 4.2.A is that

$$n(\lambda^n - \lambda) \to \left(\min_{i \in \beta} \frac{c_i}{\lambda_i}\right)\lambda \quad \text{as } n \to \infty.$$

This convergence, which partially accounts for the form of d_β in (4.10), is verified in Subsection 5.7.

Remark 10. We believe that the convergence of $\hat{D}^n_{j,h}$ still holds without the restriction $h_j = 1$.

4.4. Let us now set up the framework within which Theorem 4.1 is proved. Then some major steps in the proof will be outlined, postponing the full details to Section 5. First we formalize three space-time rescalings that will be employed. One rescaling, which has already appeared, yields diffusion limits and is denoted by a "hat". Examples are the sequences in (4.1) as well as

$$\hat{S}^n(t) = \frac{1}{n}\left[S^n(n^2t) - \mu^n n^2 t\right], \qquad \hat{R}^n(t) = \frac{1}{n}\left[R(\lfloor n^2 t\rfloor) - P' n^2 t\right],$$

$$\hat{V}_j^n(t) = \frac{1}{n\mu_j^n}\left[V_j(\lfloor n^2 t\rfloor) - n^2 t\right],$$

where S^n is the service process of the nth network, R is defined in 2.2.D and V_j in 2.2.B. The second rescaling yields fluid approximations (as in CMa; see also Subsection 4.6) and is denoted by a "bar". Specifically, consider a sequence $F^n = \{F^n(u),\ u \geq 0\}$, $n = 1, 2, \ldots$. Then, depending on the nature of the parameter u, we write $\bar{F}^n = \{\bar{F}^n(t),\ t \geq 0\}$, $n = 1, 2, \ldots$, to denote either

$$\bar{F}^n(t) = \frac{1}{n}F^n(nt)$$

when u varies continuously in $[0, \infty)$ or

$$\bar{F}^n(t) = \frac{1}{n}F^n(\lfloor nt\rfloor)$$

when u is integer-valued, for example $u = 1, 2, \ldots$. The third rescaling, denoted by double bars, has no qualitative significance and is introduced purely for notational convenience. It stands for

$$\bar{\bar{F}}^n(t) = \frac{1}{n}\bar{F}^n(nt),$$

being either $F^n(n^2t)/n^2$ or $F^n(\lfloor n^2 t\rfloor)/n^2$ as before.

4.5. The framework for proving Theorem 4.1 is built around the following two functional central limit theorems (FCLT's):

$$\hat{S}^n(t) = \bar{S}^n(nt) - \mu^n nt \to_d \hat{S}(t) \quad \text{in } D^J, \tag{4.12}$$

$$\hat{R}^n(t) = \bar{R}^n(nt) - P'nt \to_d \hat{R}(t) \quad \text{in } D^{J^2}. \tag{4.13}$$

In (4.12) which, by independence, is an immediate multivariate extension of Billingsley's (1968) FCLT for renewal processes,

$$\hat{S} = \mathrm{BM}(0, \Lambda_S) \quad \text{with} \quad \hat{S}(0) = 0 \quad \text{and} \quad (\Lambda_S)_{jk} = \delta_{jk}\mu_j b_j^2.$$

In (4.13), which is a consequence of Donsker's multivariate FCLT for partial sums, $\hat{R} = (\hat{R}^1, \ldots, \hat{R}^J)$ has independent columns that satisfy

$$\hat{R}^j = \mathrm{BM}(0, \Lambda_R^j) \quad \text{with} \quad \hat{R}^j(0) = 0 \quad \text{and} \quad (\Lambda_R^j)_{kl} = p_{jk}(\delta_{kl} - p_{jl}).$$

The random sequences $\{\hat{Q}^n(0)\}$, $\{\hat{S}^n\}$ and $\{\hat{R}^n\}$ are mutually independent. This ensures that the Brownian motions $\hat{S}$, $\hat{R}^j$, $j = 1, \ldots, J$, are mutually independent and that 4.2.B, (4.12) and (4.13) actually hold jointly. By Skorohod's representation theorem [Skorohod (1956)], there exists a single probability space supporting versions of

4.5.A $$\left(\overline{Q}^n(0), \overline{S}^n, \overline{R}^n\right), \qquad n = 1, 2, \dots$$

in which the convergence 4.2.B is almost sure and the convergence in both (4.12) and (4.13) is almost surely uniform on compact subsets of $[0, \infty)$. Consequently, we may and shall assume in the sequel that with probability 1, as $n \to \infty$,

4.5.B $$\hat{Q}^n(0) = \overline{Q}^n(0) \to \hat{Q}(0),$$

4.5.C $$\hat{S}^n(t) = \overline{S}^n(nt) - \mu^n nt \to \hat{S}(t) \quad \text{u.o.c.},$$

4.5.D $$\hat{R}^n(t) = \overline{R}^n(nt) - P'nt \to \hat{R}(t) \quad \text{u.o.c.}$$

and that the limits

4.5.E $$\hat{Q}(0), \hat{S} \text{ and } \hat{R} \text{ are independent.}$$

Furthermore, 4.5.C and 4.5.D, respectively, imply that, as $n \to \infty$,

$$\overline{S}^n(t) \to \mu t \quad \text{u.o.c.}, \tag{4.14}$$

$$\overline{R}^n(t) \to P't \quad \text{u.o.c.}, \tag{4.15}$$

hold pathwise. These last two facts follow from the following.

LEMMA 4.2. *Let $\hat{F}(t)$ and $F^n(t)$, $n = 1, 2, \dots,$ be functions in D^1 and let r and r_n, $n = 1, 2, \dots,$ be real numbers. Assume that $\hat{F}(0) = 0$ and that r is the limit of $\{r_n\}$ as $n \to \infty$. If*

$$\hat{F}^n(t) = \frac{1}{n}\left[F^n(n^2 t) - r_n n^2 t\right] \to \hat{F}(t) \quad u.o.c. \quad as\ n \to \infty, \tag{4.16}$$

then

$$\overline{F}^n(t) = \frac{1}{n} F^n(nt) \to rt \quad u.o.c. \quad as\ n \to \infty.$$

Lemma 4.2 is proved in Subsection 5.7.

4.6. As already mentioned at the end of Subsection 3.2, CMa provide FSLLN's from which fluid approximations emerge. More precisely, with only 4.5.B and (4.14)–(4.15) as assumptions, CMa show that "bar-rescaling" the queueing network by $n \to \infty$ results in convergence to the linear closed fluid network in Subsections 3.1–3.2, with initial fluid $Z(0) = \overline{Q}(0) = \hat{Q}(0)$, release capacity vector $\mu = (\mu_j)$ and $P = [p_{jk}]$. In particular, the fluid level Z, which here will be denoted by $\overline{Q}$, is derived from the almost sure limit

$$\overline{Q}^n(t) = \frac{1}{n} Q^n(nt) \to \overline{Q}(t) \quad \text{u.o.c.}, \tag{4.17}$$

and the cumulative fraction of utilized capacity $\overline{B}$, expressed in terms of the loss process Y by $\overline{B}_j(t) = (\mu_j t - Y_j(t))/\mu_j$, $t \geq 0$, is obtained from

$$(4.18)\qquad \overline{B}^n(t) = \frac{1}{n}B^n(nt) \to \overline{B}(t) \quad \text{u.o.c.}$$

Further analysis in CMa reveals that $(\overline{Q}, \overline{B})$ are piecewise-linear and that there exists a finite time τ such that

$$(4.19)\qquad \overline{Q}_\alpha(t) = 0 \quad \text{for } t \geq \tau,$$

$$(4.20)\qquad \overline{B}(t) = \overline{B}(\tau) + \rho(t-\tau), \qquad t \geq \tau.$$

The time τ is the equilibrium time mentioned at the end of Subsection 3.2. It actually equals 0 when $\overline{Q}_\alpha(0) = 0$.

4.7. The elements in 4.5.A–4.5.D will now be used as the building blocks for two representations, "bar" and "hat" that, respectively, link the fluid and diffusion approximations to the microscopic representation 2.7.A–2.7.C.

One uses first the elements in 4.5.A to construct $(\overline{Q}^n, \overline{B}^n)$ via (2.2)–(2.3), namely

$$\overline{Q}_j^n(t) = \overline{Q}_j^n(0) + \sum_{k=1}^{J} \overline{R}_j^{k,n}\left\{\overline{S}_k^n\left[\overline{B}_k^n(t)\right]\right\} - \overline{S}_j^n\left[\overline{B}_j^n(t)\right], \qquad t \geq 0,$$

$$\overline{B}_j^n(t) = \int_0^t 1\left[\overline{Q}_j^n(u) > 0\right] du, \qquad t \geq 0.$$

The bar representation is then

$$4.7.A \qquad \overline{Q}^n = \overline{X}^n + [I - P']\overline{Y}^n,$$

where

$$(4.21)\qquad \overline{X}^n(t) = \overline{Q}^n(0) + \theta^n t + \bar{\xi}^n(t),$$

$$(4.22)\qquad \theta^n = [P' - I]\mu^n,$$

$$(4.23)\qquad \begin{aligned} \bar{\xi}_j^n(t) = {} & \sum_{k=1}^{J} \left\{\overline{R}_j^{k,n}\left[\overline{S}_k^n\left(\overline{B}_k^n(t)\right)\right] - p_{kj}\overline{S}_k^n\left(\overline{B}_k^n(t)\right)\right\} \\ & + \sum_{k=1}^{J} p_{kj}\left[\overline{S}_k^n\left(\overline{B}_k^n(t)\right) - \mu_k^n \overline{B}_k^n(t)\right] \\ & - \left[\overline{S}_j^n\left(\overline{B}_j^n(t)\right) - \mu_j^n \overline{B}_j^n(t)\right], \end{aligned}$$

$$(4.24)\qquad \overline{Y}_j^n(t) = \mu_j^n\left[t - \overline{B}_j^n(t)\right],$$

$$(4.25)\qquad \overline{B}_j^n(t) = \int_0^t 1\left[\overline{Q}_j^n(u) > 0\right] du.$$

The processes $\overline{X}^n$, $\overline{Y}^n$ and $\overline{Q}^n$ uniquely satisfy 2.7.A–2.7.C (playing the roles of X, Y and Q there, respectively). From 3.3.A–3.3.C, it follows that $\overline{Y}^n = \Psi_P(\overline{X}^n)$, namely the image of $\overline{X}^n$ under oblique reflection. This allows for an easy explanation of the methodology underlying CMa: First prove the u.o.c.

convergence of $\overline{X}^n$ to some $\overline{X}$; then use the continuity 3.3.E to deduce the convergence of $\overline{Y}^n$ to $\overline{Y} = \Psi_P(\overline{X})$; finally, conclude the convergence of $\overline{Q}^n$ in 4.7.A to $\overline{Q} = \overline{X} + \Psi_P(\overline{X})$ in (4.17); the convergence of other sequences of interest readily follows. In the present paper we follow this same approach, but with "hats" instead of "bars". To this end, observe that

$$\hat{Q}^n(t) = \overline{Q}^n(nt) \quad \text{and} \quad \hat{B}^n(t) = \overline{B}^n(nt) - \rho^n nt.$$

Then define

$$\hat{Y}^n(t) = \overline{Y}^n(nt) \quad \text{and} \quad \hat{\xi}^n(t) = \bar{\xi}^n(nt),$$

to get the hat representation

$$\text{4.7.B} \qquad \hat{Q}^n = \hat{X}^n + [I - P']\hat{Y}^n.$$

Here

$$\hat{X}^n(t) = \hat{Q}^n(0) + \theta^n nt + \hat{\xi}^n(t), \tag{4.26}$$

$$\hat{\xi}^n_j(t) = \sum_{k=1}^{J} \hat{R}^{k,n}_j\Big[\overline{\overline{S}}{}^n_k\Big(\overline{\overline{B}}{}^n_k(t)\Big)\Big] + \sum_{k=1}^{J} p_{kj}\hat{S}^n_k\Big(\overline{\overline{B}}{}^n_k(t)\Big) - \hat{S}^n_j\Big(\overline{\overline{B}}{}^n_j(t)\Big), \tag{4.27}$$

$$\hat{Y}^n_j(t) = (\mu^n_j - \lambda^n_j)t - \mu^n_j\hat{B}^n_j(t), \tag{4.28}$$

$$\overline{\overline{B}}{}^n_j(t) = \frac{1}{n}\overline{B}^n_j(nt) = \int_0^t 1\big[\hat{Q}^n_j(u) > 0\big]\,du, \tag{4.29}$$

where λ^n is the inflow capacity of the network $(0, P, \mu^n)$. Like their "bar" analogs, the processes $\hat{Q}^n$ and $\hat{Y}^n$ uniquely satisfy 2.7.A–2.7.C, hence $\hat{Y}^n = \Psi_P(\hat{X}^n)$ [see, however, (4.33) for the reflection mapping that will actually be used].

Skorohod's representation 4.5.B–4.5.D will enable us to prove FCLT's for queue lengths and busy times by sample-path analysis. The workload process will be analyzed similarly, guided by (2.4) and the representations

$$\text{4.7.C} \qquad \overline{W}^n_j(t) = \frac{1}{\mu^n_j}\overline{V}^n_j\Big[\overline{Q}^n_j(t) + \overline{S}^n_j\Big(\overline{B}^n_j(t)\Big)\Big] - \overline{B}^n_j(t)$$

and

$$\text{4.7.D} \quad \hat{W}^n_j(t) = \frac{1}{\mu^n_j}\Big\{\hat{Q}^n_j(t) + \hat{V}^n_j\Big[\overline{\overline{Q}}{}^n_j(t) + \overline{\overline{S}}{}^n_j\Big(\overline{\overline{B}}{}^n_j(t)\Big)\Big] + \hat{S}^n_j\Big(\overline{\overline{B}}{}^n_j(t)\Big)\Big\}.$$

Indeed, we shall show that the sample paths of $(\hat{Q}^n, \hat{W}^n, \hat{B}^n)$, obtained from Skorohod's representation and viewed as a sequence of functions, converge u.o.c. to the appropriate limits. (Sojourn times will be analyzed probabilistically, however.) This strong convergence implies, of course, weak convergence of the original processes, as elaborated on in the introduction to Whitt (1980).

4.8. Two major steps in the proof of Theorem 4.1 are formulated as:

PROPOSITION 4.3. *Let τ be the equilibrium time for the fluid approximation, as described in* (4.19)–(4.20). *Then*

$$\hat{Q}_\alpha^n\left(t+\frac{\tau}{n}\right)=\overline{Q}_\alpha^n(nt+\tau)=\frac{1}{n}Q_\alpha^n(n^2t+n\tau)\to 0 \quad u.o.c.\ as\ n\to\infty.$$

PROPOSITION 4.4.

$$\overline{\overline{B}}^n(t)=\frac{1}{n}\overline{B}^n(nt)=\frac{1}{n^2}B^n(n^2t)\to\rho t \quad u.o.c.\ as\ n\to\infty.$$

The convergence

$$\hat{Q}_\alpha^n\to 0 \quad \text{u.o.c. in } t>0, \tag{4.30}$$

which justifies (4.2), is then a consequence of Proposition 4.3 and:

LEMMA 4.5. *Let $F(t)$ be a function in C^1 and $F^n(t)$, $n=1,2,\dots,$ a sequence in D^1. If for some $\tau\ge 0$,*

$$F^n\left(t+\frac{\tau}{n}\right)\to F(t) \quad u.o.c.\ as\ n\to\infty,$$

then

$$F^n\to F \quad u.o.c.\ in\ t>0\ as\ n\to\infty.$$

Lemma 4.5 is proved in Section 5.7. [Note the difference between (4.2) and (4.19) and the difference between Proposition 4.4 and (4.20); these are rooted in the different rescalings employed.]

The convergence of $\hat{Q}_\beta^n$ is obtained after isolating the bottleneck subnetwork. Formally, this entails writing 4.7.B in blocks α and β, then solving for $\hat{Y}_\alpha^n$ in the α block and substituting the result into the β block. These simple calculations suggest the representation

$$\hat{Q}_\beta^n=\tilde{X}^n+[I-\tilde{P}']\hat{Y}_\beta^n, \tag{4.31}$$

where

$$\tilde{X}^n=\hat{X}_\beta^n+P'_{\alpha\beta}[I-P'_\alpha]^{-1}\hat{X}_\alpha^n-P'_{\alpha\beta}[I-P'_\alpha]^{-1}\hat{Q}_\alpha^n, \tag{4.32}$$

$$\hat{Y}_\beta^n=\Psi_{\tilde{P}}(\tilde{X}^n), \tag{4.33}$$

with $\tilde{P}$ given in (4.6) and $\hat{X}$ in (4.26). By analogy to the "bar" convergence in 4.7.A, one would now anticipate u.o.c. convergence of $\tilde{X}^n$ to X in (4.4), thus implying (4.3)–(4.7) in view of the continuity of $\Psi_{\tilde{P}}$. But two obstacles arise: first, the sequence $\tilde{X}^n$ need not converge u.o.c. and second, $\tilde{X}^n$ does not necessarily belong to the domain of $\Psi_{\tilde{P}}$. The first obstacle prevails when a nonnegligible fraction of customers initially occupy the nonbottlenecks: $\hat{Q}_\alpha^n(0)\to\hat{Q}_\alpha(0)\neq 0$, or equivalently, the u.o.c. convergence in (4.30) does not extend to $t \geq 0$ (see also Remarks 3 and 4 in Subsection 4.3). The second obstacle is encountered when all the customers happen to accumulate at the nonbottleneck stations for a long enough period: $\hat{Q}^n_\beta = 0$ during some time interval. The two obstacles will be circumvented as follows.

First we shift, via 3.5.C, the time origin of the nth network from $t = 0$ to time $n\tau$ [see (5.1)–(5.3)]. As apparent from Proposition 4.3, such a shift is negligible in the diffusion time scale and the shifted networks do satisfy (4.30) u.o.c. In other words, at the end of the negligible period $[0, n\tau]$, n large, essentially all the customers will occupy only the bottlenecks. Next we verify that only a negligible fraction of the customers actually returns to the nonbottlenecks during the time period $[n\tau, n\tau + n^2 T_n)$, for some specified sequence $T_n \to \infty$ as $n \to \infty$ [see (5.18)]. This renders (4.33) valid for the nth shifted network at least over $t \in [0, T_n)$, $T_n \to \infty$. From 3.3.G, it will then follow that $\hat{Y}^n_\beta \to Y$ in (4.7), $\hat{Q}^n_\beta \to \hat{Q}_\beta$ in $t > 0$ will then be deduced and (4.3)–(4.7) will finally be established.

Propositions 4.3 and 4.4, which constitute the main hurdle in establishing Theorem 4.1, are verified in Subsections 5.1–5.3. The convergence of $\hat{Q}^n_\beta$ and $\hat{B}^n$ is deduced in Subsection 5.4 [the latter follows from the convergence of $\hat{Y}^n$, in view of (4.28)]. The limits (4.8) and (4.9) for workloads and sojourn times are established in Subsections 5.5 and 5.6, respectively, and their proofs are self-contained.

4.9. Readers may wonder why the valid relation $\hat{Y}^n = \Psi_P(\hat{X}^n)$ is not pursued directly. The reason is that the drift of $\hat{X}^n$ in (4.26), hence $\hat{X}^n$ itself, need *not* converge, which is manifested by the divergence of $\hat{Y}^n_\alpha$ to infinity. Indeed, from Proposition 4.4 and (4.24) follows that $\overline{\overline{Y}}^n(t) \to (\mu - \lambda)t$ and consequently $\hat{Y}^n_\alpha = n\overline{\overline{Y}}^n_\alpha$ blows up.

A related observation is that, while $\hat{X}^n$ need not converge, the *sum* of the first two terms on the right-hand side of (4.32) does converge u.o.c. This will follow from 5.3.A and the convergence

$$n\left\{\theta^n_\beta + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\theta^n_\alpha\right\} \to [\tilde{P}' - I]c_\beta$$

as $n \to \infty$, which is established prior to (5.19). [The third term in (4.32) can be guaranteed to converge u.o.c. only in $t > 0$.]

5. Proofs of the limit theorems for irreducible closed networks. The present section is entirely devoted to proving the results exposited in Section 4. Many arguments are based on facts established in CMa, so readers interested in details probably will find it necessary to consult it frequently.

5.1. *Proofs of Propositions* 4.3 *and* 4.4. We shall prove below the following two implications:

5.1.A If $\overline{\overline{B}}^n(t) = \overline{B}^n(nt)/n$ converges u.o.c. along a subsequence, then $\hat{Q}^n_\alpha(t + \tau/n)$ also converges u.o.c. to 0 along this subsequence, or in other words, Proposition 4.3 holds for that subsequence as well.

5.1.B If along a subsequence both Proposition 4.3 holds and $\overline{\overline{B}}^n$ converges u.o.c., then the latter convergence is to ρt, or in other words, Proposition 4.4 holds for that subsequence as well.

A consequence of 5.1.A and 5.1.B is that any u.o.c. convergent subsequence of $\overline{\overline{B}}^n$ must converge to ρt. Now (4.29) implies that $\overline{\overline{B}}^n$ is uniformly Lipshitz. By Arzela–Ascoli's theorem, any subsequence of $\overline{\overline{B}}^n$ has a convergent subsequence which, in turn, converges to ρt u.o.c. This establishes Proposition 4.4 and then Proposition 4.3 [the latter in view of 5.1.A].

In obtaining both 5.1.A and 5.1.B, one uses the fact that for $\hat{\xi}^n$ in (4.27):

5.1.C If $\overline{\overline{B}}^n$ converges u.o.c. along a subsequence, then $\hat{\xi}^n$ converges u.o.c. along that subsequence to a *continuous* limit which vanishes at $t = 0$.

To prove 5.1.C, first observe that $\overline{\overline{S}}^n(t) \to \mu t$ u.o.c. by 4.2.A and 4.5.C, and that the u.o.c. limit points of $\overline{\overline{B}}^n$ must be continuous. Then note that the u.o.c. limits of $\hat{S}^n$ and $\hat{R}^n$ in 4.5.C–4.5.D are also continuous, being Brownian sample paths. Finally, apply the deterministic time-change theorem [Whitt (1980), Theorem 3.1] which, for convenience and future reference, is stated here as:

5.1.D Let $\{F_n, n \geq 1\}$ and $\{c_n, n \geq 1\}$ be sequences in D^J. Assume that c_n is nondecreasing with $c_n(0) = 0$. If $(F_n(t), c_n(t))$ converges u.o.c. to a continuous pair $(F(t), c(t))$, then $F_n(c_n(t))$ converges u.o.c. to $F(c(t))$.

The verifications of 5.1.A and 5.1.B remain the missing link in the proofs of Propositions 4.3–4.4. We establish the first in the next subsection and the second in Subsection 5.3. For both proofs, the subsequence in question is taken to be the sequence itself for ease of notation.

5.2 Proof of 5.1.A. The proof will be carried out as though $\hat{Q}^n_\alpha(0) \to \hat{Q}_\alpha(0) = 0$, hence $\tau = 0$ [see the comment following (4.20)]. For otherwise, we start the nth network at time $n\tau$ by letting

$$\vec{X}^n(t) = \hat{Q}^n(\tau/n) + \hat{X}^n(t + \tau/n) - \hat{X}^n(\tau/n), \tag{5.1}$$

$$\vec{Y}^n(t) = \hat{Y}^n(t + \tau/n) - \hat{Y}^n(\tau/n). \tag{5.2}$$

Then property 3.5.C reads

$$\vec{Q}^n(t) = \hat{Q}^n(t + \tau/n) = \vec{X}^n(t) + [I - P']\vec{Y}^n(t), \qquad t \geq 0. \tag{5.3}$$

Now the process $\vec{X}^n$ has the form

$$\begin{aligned} \vec{X}^n(t) &= \vec{Q}^n(0) + \vec{\xi}^n(t) + \theta^n nt \quad \text{with} \\ \vec{\xi}^n(t) &= \hat{\xi}^n(t + \tau/n) - \hat{\xi}^n(\tau/n). \end{aligned} \tag{5.4}$$

By (4.17) and (4.19), $\vec{Q}^n_\alpha(0) = \hat{Q}^n_\alpha(\tau/n) = \overline{Q}^n_\alpha(\tau) \to 0$ and 5.1.C implies that $\vec{\xi}^n(t)$ converges u.o.c. to a continuous limit. These last two facts will be seen to guarantee the u.o.c. convergence $\vec{Q}_\alpha^n \to 0$, thus reducing 5.1.A to the case $\tau = 0$ as claimed.

Consider the subset of stations $G = \{j\colon \theta_j < 0\}$, where $\theta = [P' - I]\mu$ is the limit of θ^n in (4.22). Stations in G are the subcritical stations in CMa. It is shown there, in Lemma 4.1, that $G = \varnothing$ implies $\alpha = \varnothing$, in which case there is nothing to prove. Let us assume, therefore, that $G \neq \varnothing$ and verify that:

5.2.A For $j \in G$, $\hat{Q}_j^n(t) \to 0$ u.o.c.

To this end, set $\varepsilon_n = \max_{j \in \alpha}\{\overline{Q}_j^n(0)\}$ and note that $\varepsilon_n \to 0$. Then introduce for $t \geq 0$,

$$\nu_j^n(t) = \sup\left\{s \leq t\colon \hat{Q}_j^n(s) \leq \varepsilon_n\right\},$$

[abbreviated as ν_j^n when convenient and well defined because the set over which the supremum is taken always contains $s = 0$: $\hat{Q}_j^n(0) = \overline{Q}_j^n(0) \leq \epsilon_n$]. Now $\hat{Q}_j^n$ is an RCLL step function. It follows from the definition of $\nu_j^n(t)$ that $\hat{Q}_j^n(\nu_j^n -) \leq \varepsilon_n$. Furthermore, if $\nu_j^n(t) < t$, then $\hat{Q}_j^n(s) > \varepsilon_n \geq 0$ for $s \in [\nu_j^n(t), t]$. Using the complementarity condition 2.7.C, applied to the hat representation 4.7.B, implies that $\hat{Y}_j^n(\nu_j^n(t)) = \hat{Y}_j^n(t)$, which also holds when $\nu_j^n(t) = t$. One utilizes all this in

$$\begin{aligned}
(5.5)\qquad -\varepsilon_n &\leq \hat{Q}_j^n(t) - \hat{Q}_j^n(\nu_j^n -)\\
&= \hat{\xi}_j^n(t) - \hat{\xi}_j^n(\nu_j^n -) + n\theta_j^n(t - \nu_j^n) - \sum_{k=1}^{J} p_{kj}\left[\hat{Y}_k^n(t) - \hat{Y}_k^n(\nu_j^n)\right]\\
&\leq \hat{\xi}_j^n(t) - \hat{\xi}_j^n(\nu_j^n -) + n\theta_j^n(t - \nu_j^n).
\end{aligned}$$

Consequently, for $t \geq 0$,

$$(5.6)\qquad 0 \leq (-\theta_j^n)\left[t - \nu_j^n(t)\right] \leq \frac{\varepsilon_n}{n} + \frac{1}{n}\hat{\xi}_j^n(t) - \frac{1}{n}\hat{\xi}_j^n(\nu_j^n -).$$

The convergence of $\hat{\xi}^n$ in 5.1.C guarantees that both $\hat{\xi}_j^n(t)$ and $\hat{\xi}_j^n(\nu_j^n(t) -)$ are uniformly bounded on any compact subset of $[0, \infty)$. One concludes from (5.6) and $-\theta_j^n \to -\theta_j > 0$, $j \in G$, that

$$\nu_j^n(t) \to t \quad \text{u.o.c.}$$

Going back to (5.5), we have

$$0 \leq \hat{Q}_j^n(t) \leq \hat{\xi}_j^n(t) - \hat{\xi}_j^n(\nu_j^n -) + \hat{Q}_j^n(\nu_j^n -),$$

for all $t \geq 0$. Finally, $0 \leq \hat{Q}_j^n(\nu_j^n -) \leq \varepsilon_n$, combined with 5.1.C–5.1.D, establishes 5.1.A for stations in G.

The extension to all $j \in \alpha$ is achieved through an elimination procedure similar to the one used to prove Theorem 5.2 in CMa. Indeed, recall (4.26) to rewrite 4.7.B in block form as

$$(5.7)\qquad \hat{Q}_G^n(t) = \hat{Q}_G^n(0) + n\theta_G^n t + \hat{\xi}_G^n(t) - P'_{HG}\hat{Y}_H^n(t) + [I - P'_G]\hat{Y}_G^n(t), \qquad t \geq 0,$$

$$(5.8)\qquad \hat{Q}_H^n(t) = \hat{Q}_H^n(0) + n\theta_H^n t + \hat{\xi}_H^n(t) - P'_{GH}\hat{Y}_G^n(t) + [I - P'_H]\hat{Y}_H^n(t), \qquad t \geq 0,$$

where H is the complement of G. Stations in G are nonbottlenecks, hence $[I - P'_G]$ is invertible by Lemma 4.4 in CMa. Isolating $\hat{Y}_G^n$ in (5.7) and substituting the outcome, together with (4.26), in (5.8) gives

$$(5.9)\qquad \hat{Q}_H^n = \tilde{X}_H^n + \left[I - \tilde{P}'_H\right]\hat{Y}_H^n,$$

where

$$(5.10)\qquad \tilde{X}_H^n(t) = \tilde{Q}_H^n(0) + n\tilde{\theta}_H^n t + \tilde{\xi}_H^n(t),$$

$$(5.11)\qquad \tilde{Q}_H^n(0) = \hat{Q}_H^n(0) + P'_{GH}[I - P'_G]^{-1}\hat{Q}_G^n(0),$$

$$(5.12)\qquad \tilde{\theta}_H^n = \theta_H^n + P'_{GH}[I - P'_G]^{-1}\theta_G^n = \left[\tilde{P}'_H - I\right]\mu_H^n,$$

$$(5.13)\qquad \tilde{\xi}_H^n(t) = \hat{\xi}_H^n(t) + P'_{GH}[I - P'_G]^{-1}\hat{\xi}_G^n(t) - P'_{GH}[I - P'_G]^{-1}\hat{Q}_G^n(t),$$

$$(5.14)\qquad \tilde{P}_H = P_H + P_{HG}[I - P_G]^{-1}P_{GH}.$$

The representation (5.9), like its predecessor 4.7.B, conforms to 2.7.A–2.7.C, but with a switching matrix $\tilde{P}_H$. The triplet $(0, \tilde{P}_H, \mu_H)$ is the image of $(0, P, \mu)$ through the transformation (4.4) in Subsection 4.6 of CMa. Since this transformation leaves all the network's characteristics derived from $(0, P, \mu)$ intact (CMa, Lemmas 4.3 and 4.5), we can now define $\tilde{G} = \{j \in H\colon \tilde{\theta}_j < 0\}$, where $\tilde{\theta}_H = (\tilde{P}'_H - I)\mu_H$, then restart the process, but confined to the stations in H. Formally, this entails proving 5.2.A for $j \in \tilde{G}$, through (5.9)–(5.14), after substituting $\tilde{P}_H$ for P, $\tilde{X}_H^n$ for $\hat{X}^n$, $\tilde{\theta}_H^n$ for θ^n and $\tilde{\theta}_H$ for θ. We have now established 5.2.B, or equivalently 5.1.A, for stations in G and $\tilde{G}$. Repeating the process if necessary, each time 5.2.B is reproved verifies 5.1.A for at least one more station in α. By Lemmas 4.5 and 4.1, both in CMa, within $|\alpha|$ steps at the most only bottlenecks remain and the proof of 5.1.A then ends. □

5.3 PROOF OF 5.1.B. The conclusions in both 5.1.A and 5.1.C prevail. In particular, 5.1.C can be written as:

5.3.A $\hat{\xi}^n$ converges u.o.c. to a limit $\hat{\xi}$ which is continuous with $\hat{\xi}(0) = 0$.

As a first step we verify that $\hat{Y}_\beta^n$ converges u.o.c., then identify its limit [in 5.3.C]. Consider the shifted network in (5.3), write it in α and β blocks as in (5.7)–(5.8), then derive the analog of (5.9)–(5.14) in the form

$$(5.15)\qquad \vec{Q}_\beta^n = \tilde{X}_\beta^n + \left[I - \tilde{P}'_\beta\right]\vec{Y}_\beta^n,$$

where

$$\tilde{X}^n_\beta(t) = \tilde{Q}^n_\beta(0) + n\tilde{\theta}^n_\beta t + \tilde{\xi}^n_\beta(t) \tag{5.16}$$

$$\tilde{Q}^n_\beta(0) = \vec{Q}^n_\beta(0) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\vec{Q}^n_\alpha(0)$$

$$\tilde{\theta}^n_\beta = \left[\tilde{P}'_\beta - I\right]\mu^n_\beta$$

$$\tilde{\xi}^n_\beta(t) = \vec{\xi}^n_\beta(t) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\left[\vec{\xi}^n_\alpha(t) - \vec{Q}^n_\alpha(t)\right] \tag{5.17}$$

$$\vec{Y}^n_\beta = \Psi_{\tilde{P}_\beta}\left(\tilde{X}^n_\beta\right) \quad \text{on } [0, T_n), \tag{5.18}$$

$$T_n = \min\left\{t \geq 0\colon e'\tilde{X}^n_\beta(t) = 0\right\}.$$

[The matrix $\tilde{P}_\beta$ coincides with $\tilde{P}$ in (4.6).] We now verify the u.o.c. convergence of $\tilde{X}^n_\beta$ in (5.16) and then justify, via $T_n \to \infty$, an application of 3.3.G to (5.18).

The convergence of $\tilde{Q}^n_\beta(0)$ holds because $\vec{Q}^n_\alpha(0) = \overline{Q}^n_\alpha(\tau) \to 0$ by (4.19) and also

$$\vec{Q}^n_\beta(0) = \overline{Q}^n_\beta(\tau) \to \hat{Q}_\beta(0) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{Q}_\alpha(0),$$

from Theorem 7.1 in CMa. Now Corollary 4.6 in CMa says that $[\tilde{P}'_\beta - I]\mu_\beta = 0$, so

$$n\tilde{\theta}^n_\beta = \left[\tilde{P}'_\beta - I\right]\left[n\left(\mu^n_\beta - \mu_\beta\right)\right] \to \left[\tilde{P}'_\beta - I\right]c_\beta \quad \text{as } n \to \infty,$$

in view of 4.2.A. As already noted after (5.4), 5.3.A implies that $\vec{\xi}^n \to \hat{\xi}$ u.o.c. Together with Proposition 4.3, this guarantees the u.o.c. convergence of $\tilde{\xi}^n_\beta$ in (5.17), as well as

$$\tilde{X}^n_\beta(t) \to X(t) \quad \text{u.o.c.},$$

where

$$X(t) = X(0) + \left[\tilde{P}'_\beta - I\right]c_\beta + \hat{\xi}_\beta(t) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{\xi}_\alpha(t), \tag{5.19}$$

$$X(0) = \hat{Q}_\beta(0) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{Q}_\alpha(0). \tag{5.20}$$

Since the matrix $\tilde{P}_\beta$ is stochastic, multiplying (5.15) by e' yields

$$e'\tilde{X}^n_\beta(t) = e'\vec{Q}^n_\beta(t) = e'\vec{Q}^n(t) - e'\vec{Q}^n_\alpha(t) = 1 - e'\vec{Q}^n_\alpha(t), \qquad t \geq 0,$$

which converges u.o.c. to $1 - e'\vec{Q}_\alpha(t) = 1$ in view of Proposition 4.3. This convergence implies that $T_n \to \infty$ for T_n in (5.18) and that $e'X(t) = 1$ for all $t \geq 0$, thus $X \in C^{|\beta|}_{>}$. Letting $n \to \infty$ in (5.18), one concludes from the continuity 3.3.G of the oblique reflection mapping that:

5.3.B $\vec{Y}^n_\beta \to Y$ u.o.c., where $Y = \Psi_{\tilde{P}_\beta}(X)$ and X is given by (5.19)–(5.20).

It is also demonstrated in Theorem 7.1 of CMa that $\hat{Y}^n_\beta(\tau/n) = \overline{Y}^n_\beta(\tau) \to \overline{Y}_\beta(\tau) = 0$. From (5.2) and 5.3.B it follows that $\hat{Y}^n_\beta(t + \tau/n) \to Y(t)$ u.o.c., hence $\hat{Y}^n_\beta(t) \to Y(t)$ for all $t > 0$ by Lemma 4.5. The latter convergence also holds at $t = 0$, since the sequence and its limit vanish there. We now notice that $\hat{Y}_\beta^n$ and $Y(t)$ are both nondecreasing and continuous, and pointwise convergence in $t \geq 0$ of such functions implies u.o.c. convergence [Resnick (1987), Chapter 1, or the remark after the proof of Theorem 6.1 in Whitt (1980)]. Thus, in general:

5.3.C $\hat{Y}_\beta^n(t) \to Y(t)$ u.o.c., where $Y = \Psi_{\hat{P}_\beta}(X)$ and X is given by (5.19)–(5.20).

Finally, we are ready to prove 5.1.B, aided by (4.24). From 5.3.C we have $\bar{\bar{Y}}_\beta^n \to 0$ u.o.c., or

$$\bar{\bar{B}}_\beta^n(t) = \frac{1}{n}\bar{B}_\beta^n(nt) \to et = \rho_\beta t \quad \text{u.o.c.},$$

since $\lambda_j = \mu_j$ for bottlenecks $j \in \beta$. This establishes the β part of 5.1.B. For the α part, note first that $\bar{\bar{\xi}}^n(t) \to 0$, because $\hat{\xi}^n$ converges [see 5.3.A]. The coordinates of $Q^n(n^2 t)$ are nonnegative and sum up to n, hence also $\bar{\bar{Q}}^n \to 0$ u.o.c. Substituting (4.22) into 4.7.B, writing it in blocks α and β and rearranging terms yields for $t \geq 0$,

$$\begin{aligned}\hat{Y}_\alpha^n(t) - \mu_\alpha^n nt = {}& [I - P'_\alpha]^{-1}\big[\hat{Q}_\alpha^n(t) - \hat{Q}_\alpha^n(0) - \hat{\xi}_\alpha^n(t) + P'_{\beta\alpha}\hat{Y}_\beta^n(t)\big] \\ & -[I - P'_\alpha]^{-1} P'_{\beta\alpha}\mu_\beta^n nt,\end{aligned}$$

which becomes

$$\begin{aligned}(5.21)\qquad & \bar{\bar{Y}}_\alpha^n(t) - (\mu_\alpha^n - \lambda_\alpha^n)t \\ & = [I - P'_\alpha]^{-1}\big[\bar{\bar{Q}}_\alpha^n(t) - \bar{\bar{Q}}_\alpha^n(0) - \bar{\bar{\xi}}_\alpha^n(t) + P'_{\beta\alpha}\bar{\bar{Y}}_\beta^n(t)\big] \\ & \quad +[I - P'_\alpha]^{-1}P'_{\beta\alpha}(\lambda_\beta^n - \lambda_\beta)t - [I - P'_\alpha]^{-1}P'_{\beta\alpha}(\mu_\beta^n - \mu_\beta)t,\end{aligned}$$

due to $\mu_\beta = \lambda_\beta$ and $\lambda_\alpha^n = [I - P'_\alpha]^{-1}P'_{\beta\alpha}\lambda_\beta^n$ (the latter in view of $\lambda^n = P'\lambda^n$). Now recall the u.o.c. convergence to 0 of $\bar{\bar{Q}}_\alpha^n$, $\bar{\bar{Y}}_\beta^n$ and $\bar{\bar{\xi}}_\alpha^n$ and the convergence of $(\lambda^n - \lambda)$ and $(\mu^n - \mu)$ [by Remark (9) in Subsection 4.3 and by 4.2.A, respectively]. These guarantee that $\bar{\bar{Y}}_\alpha^n(t)$ converges to $(\mu_\alpha - \lambda_\alpha)t$ u.o.c. Equivalently,

$$\bar{\bar{B}}_\alpha^n(t) = \frac{1}{n}\bar{B}_\alpha^n(nt) \to \rho_\alpha t \quad \text{u.o.c.},$$

which completes the proof of 5.1.B. □

5.4. *Finalizing the convergence of* $\hat{Q}^n$ *and* $\hat{B}^n$. The arguments in Subsections 5.2–5.3 were actually carried out along a subsequence. This is of no concern now, since Propositions 4.3–4.4 have been verified. Aided by the time change theorem 5.1.D, we first identify the u.o.c. limit $\hat{\xi}$ of $\hat{\xi}^n$ [in (4.27); see also 5.3.A] as

$$(5.22)\qquad \hat{\xi}_j(t) = \sum_{k=1}^{J}\hat{R}_j^k(\lambda_k t) + \sum_{k=1}^{J} p_{kj}\hat{S}_k(\rho_k t) - \hat{S}_j(\rho_j t).$$

The convergence

$$\hat{Q}_\alpha^n \to \hat{Q}_\alpha = 0 \quad \text{u.o.c. in } t > 0 \quad \text{as } n \to \infty$$

was justified in (4.30). Replaying Subsection 5.3 now proves 5.3.C along the whole sequence, thus $\vec{Q}^n_\beta(t) \to \hat{Q}_\beta(t)$ u.o.c. and the limit is identified from (5.15), (5.19) and (5.20) as $\hat{Q}_\beta = X + [I - \tilde{P}'_\beta]Y$, $t \geq 0$. Now observe that the limit is continuous, so Lemma 4.5 guarantees that

$$\hat{Q}^n_\beta(t) \to \hat{Q}_\beta(t) \quad \text{u.o.c. in } t > 0 \quad \text{as } n \to \infty.$$

Since $\tilde{P}_\beta = \tilde{P}$ in (4.6) and $Y = \Psi_{\tilde{P}_\beta}(X)$, we have established (4.3)–(4.7).

Now multiply (5.21) by n and conclude from Remark 9 in Subsection 4.3 that $[\hat{Y}^n_\alpha(t) - (\mu_\alpha - \lambda_\alpha)nt]$ converges u.o.c. in $t > 0$ to

$$-[I - P'_\alpha]^{-1}\left[\hat{Q}_\alpha(0) + \hat{\xi}_\alpha(t) - P'_{\beta\alpha}Y_\beta(t) + P'_{\beta\alpha}\, d_\beta t\right],$$

where

$$d_\beta = c_\beta - \left(\min_{i \in \beta} \frac{c_i}{\lambda_i}\right)\lambda_\beta.$$

The relation

$$\hat{B}^n(t) = -\operatorname{diag}\left[(\mu^n)^{-1}\right]\left[\hat{Y}^n(t) - (\mu^n - \lambda^n)nt\right].$$

finalizes the analysis of queue lengths and busy times.

5.5. *The convergence of* $\hat{W}^n$. Consider the right-hand side of 4.7.D. Its first term converges u.o.c. to $\hat{Q}_j(t)/\mu_j$. The third term converges u.o.c. to $\hat{S}(\rho_j t)/\mu_j$ in view of Proposition 4.4, the convergence in 4.2.A and 4.5.C, and the time-change theorem 5.1.D. To identify the limit of the second term notice, as in Theorem 3 and Remark (3.3) in Glynn and Whitt (1986), that the convergence 4.5.C is pathwise equivalent to

$$\text{5.5.A} \qquad \hat{V}^n_j(t) = \frac{1}{\mu^n_j}\left[\bar{V}^n_j(nt) - nt\right] \to \hat{V}_j(t) \quad \text{u.o.c. as } n \to \infty,$$

where

$$\hat{V}_j(t) = -\frac{1}{\mu_j}\hat{S}_j\left(\frac{t}{\mu_j}\right), \qquad j = 1, \dots, J.$$

Proposition 4.4, 4.5.C and 5.1.D now imply that

$$\bar{\bar{Q}}^n_j(t) + \bar{\bar{S}}^n_j\left(\bar{\bar{B}}^n_j(t)\right) \to 0 + \mu_j \rho_j t = \lambda_j t \quad \text{u.o.c. as } n \to \infty.$$

Finally, use 5.5.A and 5.1.D to verify that the second term in 4.7.D converges to $-\hat{S}(\rho_j t)/\mu_j$ which, in turn, cancels out the limit of the third term. This establishes (4.8).

5.6. *The convergence of* $\hat{D}^n$. The proof is probabilistic and is based on Section 5 in Reiman (1984). Fix a station j and j-accessible h throughout the discussion and recall the representation (2.5):

$$\text{(5.23)} \qquad D^n_{j,h}(t) = \sum_{k=1}^{J} \sum_{i=1}^{h_k} W^n_k\left(\tau^n_{j,h,k,i}(t)\right).$$

By the weak convergence of $\hat{W}^n$ and the random time-change theorem [Billingsley (1968), page 143], it suffices to prove that as $n \to \infty$,

$$\bar{\bar{\tau}}^n_{j,h,k,i}(t) = \tau^n_{j,h,k,i}(n^2 t)/n^2 \to_p t. \tag{5.24}$$

(Before proceeding, readers may find it helpful to review our conventions for $\to_p$ and $\to_d$ convergence of stochastic processes; they are introduced in the Convergence part of Subsection 1.6.) From the definition of $\tau^n_{j,h,k,i}(t)$,

$$t \le \bar{\bar{\tau}}^n_{j,h,k,i}(t) \le \bar{\bar{\tau}}^n_{j,h}(t) + \bar{\bar{D}}^n_{j,h}(t), \tag{5.25}$$

where $\tau^n_{j,h}(u) = \tau^n_{j,h,j,1}(u)$ is the arrival time to station j of the first customer to follow h upon entering j after time u. Thus, (5.25) will imply (5.24) once we show that, as $n \to \infty$,

5.6.A $$\bar{\bar{\tau}}^n_{j,h}(t) \to_p t,$$

5.6.B $$\bar{\bar{D}}^n_{j,h}(t) \to_p 0.$$

Proof of 5.6.A. Denote by

$$L^n_j(t) = \sum_{k=1}^{J} R^{k,n}_j\{S^n_k[B^n_k(t)]\}$$

the total number of (endogenous) arrivals to station j up to time t. First, we claim that

$$\bar{\bar{L}}^n_j(t) \to_d \lambda_j t. \tag{5.26}$$

Indeed, Proposition 4.4 guarantees the weak convergence $\bar{\bar{B}}^n(t) \to_d \rho t$ and (4.12)–(4.13) imply that $\bar{\bar{S}}^n(t) \to_d \mu t$ and $\bar{\bar{R}}^n(t) \to_d P't$. Thrice using the time-change theorem, one deduces (5.26) from the relation $\sum_{k=1}^{J} \lambda_k p_{kj} = \lambda_j$.

Let us remark that the event $\tau^n_{j,h}(n^2 t) \ge n^2(t + \varepsilon)$ indicates that none of the customers who arrive at j during $[n^2 t, n^2(t+\varepsilon))$ actually follows h upon entering j. We now have

$$P\Big\{\sup_{0 \le u \le t} \big[\bar{\bar{\tau}}^n_{j,h}(u) - u\big] > \varepsilon\Big\}$$
$$\le P\Big\{\bar{\bar{L}}^n_j(t+1) \ge 2\lambda_j(t+1)\Big\} \tag{5.27}$$
$$+P\Big\{\inf_{0 \le u \le t} \Big[\bar{\bar{L}}^n_j(u+\varepsilon) - \bar{\bar{L}}^n_j(u)\Big] \le \tfrac{1}{2}\lambda_j \varepsilon\Big\} \tag{5.28}$$
$$+P\{E^n_t\}, \tag{5.29}$$

where E^n_t is the intersection of the following two events: (i) the total number of arrivals to j during $[0, n^2(t+1)]$ is less than $2\lambda_j(t+1)n^2$; (ii) for some $u \in [0,t]$, the number of arrivals to j during $[n^2 u, n^2(u+\varepsilon))$ exceeds $\frac{1}{2}\lambda_j \varepsilon n^2$ and none of the customers who perform these arrivals follows h upon entering j. In view of (5.26), the terms in (5.27)–(5.28) converge to zero for every $t \ge 0$

as $n \to \infty$. Following (44) in Reiman (1984), we now majorize (5.29) by a sequence that also converges to zero. To this end, let H_i stand for the event that the customer who performs the ith arrival to j (starting the count at time $t = 0$) does follow h upon entering j. Then there exists some $p > 0$ such that $P\{H_i\} = p$ for all $i \geq 1$, because h is j-accessible. Now recall that a return to j is allowed only as the last visit on the route associated with h ($h_j = 1$). This guarantees that $H_1, H_2, \dots$ are independent. Simplifying notations with $c_1 = 2\lambda_j(t + 1)$ and $c_2 = \frac{1}{2}\lambda_j \varepsilon$, we now have

$$\begin{aligned} P\{E_t^n\} &\leq P\left\{ \bigcup_{l=1}^{\lfloor c_1 n^2 \rfloor} \bigcap_{i=l}^{l+\lfloor c_2 n^2 \rfloor} H_i^c \right\} \\ &\leq \sum_{l=1}^{\lfloor c_1 n^2 \rfloor} P\left\{ \bigcap_{i=l}^{l+\lfloor c_2 n^2 \rfloor} H_i^c \right\} \\ &= \lfloor c_1 n^2 \rfloor (1 - p)^{\lfloor c_2 n^2 \rfloor + 1} \to 0 \quad \text{as } n \to \infty. \end{aligned}$$

The proof of 5.6.A is now complete. □

Proof of 5.6.B. The event $D_{jh}^n(n^2 t) > n^2 \varepsilon$ implies that there exists a station along h at which our customer experiences a wait no less than $n^2\varepsilon/(e'h)$. Formally, $W_k^n[\tau_{j,h,k,i}^n(n^2 t)] \geq n^2\varepsilon/(e'h)$ for some pair k, i. Suppose this happens for the first time at the ith visit to station k. Then this visit starts at time $\tau_{j,h,k,i}^n(n^2 t)$ which must satisfy $\tau_{j,h,k,i}^n(n^2 t) \leq \tau_{j,h}^n(n^2 t) + n^2 \varepsilon$. Consequently, for $0 < a < b$, we have

$$\begin{aligned} & P\left\{ \sup_{a \leq t \leq b} \overline{\overline{\overline{D}}}{}_{j,h}^n(t) > \varepsilon \right\} \\ & \leq P\left\{ \sup_{a \leq t \leq b} \max_{k,i} W_k^n \left[\tau_{j,h,k,i}^n(n^2 t) \right] \geq n^2 \varepsilon/(e'h) \right\} \\ (5.30) \qquad & \leq P\left\{ \sup_{a \leq t \leq b} \left| \overline{\overline{\overline{\tau}}}{}_{j,h}^n(t) - t \right| > \varepsilon \right\} + \sum_{k=1}^{J} P\left\{ \sup_{a \leq u \leq b + 2\varepsilon} \overline{\overline{\overline{W}}}{}_k^n(u) \geq \varepsilon/e'h \right\}. \end{aligned}$$

The first term in (5.30) converges to zero by 5.6.A. The limit of the second term vanishes because of the weak convergence of $\hat{W}^n$. The proof is now complete. □

5.7. We conclude Section 5 with proofs of Lemma 4.2, Lemma 4.5 and Remark 9 from Subsection 4.3.

Proof of Lemma 4.2. Use (4.16) and the time-change theorem 5.1.D with elements $F_n = \hat{F}^n$, $F = \hat{F}$, $\tau_n(t) = t/n$ and $\tau(t) = 0$ to conclude that

$$\hat{F}^n(t/n) = \frac{1}{n} F^n(nt) - r_n t \to \hat{F}(0).$$

Lemma 4.2 now follows from the assumptions that $\hat{F}(0) = 0$ and $r_n \to r$.

(Whitt suggested this proof of Lemma 4.2, thus trivializing our original version.) □

PROOF OF LEMMA 4.5. Choose an interval $[a, b]$ with $a > 0$. Then for all $n \geq N$, N large enough, we have

$$\begin{aligned}
\sup_{a \leq u \leq b} |F^n(u) - F(u)| &\leq \sup_{\tau/n \leq u \leq N + \tau/n} |F^n(u) - F(u)| \\
&= \sup_{0 \leq t \leq N} |F^n(t + \tau/n) - F(t + \tau/n)| \\
&\leq \sup_{0 \leq t \leq N} |F^n(t + \tau/n) - F(t)| + \sup_{0 \leq t \leq N} |F(t + \tau/n) - F(t)|. \qquad (5.31)
\end{aligned}$$

Now the first summand in (5.31) converges to 0 by assumption and the second by the continuity of F. □

PROOF OF REMARK 9 FROM SUBSECTION 4.3. Since P is irreducible and λ^n and λ are both positive eigenvectors of P', there exist positive constants r_n for which $\lambda^n = r_n \lambda$ (Lemma 3.3 in CMa). From 4.2.A, $\mu^n = \mu + c^n/n$ with $c^n \to c$. Now recall the definitions of λ^n, λ and β and deduce that

$$\max_i \frac{\lambda_i^n}{\mu_i^n} = 1,$$

$$\frac{\mu_i}{\lambda_i} = 1 \quad \text{for } i \in \beta \quad \text{and} \quad \frac{\mu_i}{\lambda_i} > 1 \quad \text{for } i \notin \beta.$$

Consequently, for n large enough,

$$r_n = \min_i \frac{\mu_i^n}{\lambda_i} = \min_i \left\{ \frac{\mu_i}{\lambda_i} + \frac{1}{n} \frac{c_i^n}{\lambda_i} \right\} = 1 + \frac{1}{n} \min_{i \in \beta} \frac{c_i^n}{\lambda_i}.$$

One concludes that

$$n(\lambda^n - \lambda) = n(r_n - 1)\lambda = \left(\min_{i \in \beta} \frac{c_i^n}{\lambda_i} \right) \lambda \to \left(\min_{i \in \beta} \frac{c_i}{\lambda_i} \right) \lambda$$

as $n \to \infty$. □

6. Diffusion approximations for open networks.

The present section is devoted to a presentation of a FCLT which is the analog for open networks of Theorem 4.1. The key ideas behind the analysis of both closed and open networks are similar, hence many details are omitted. Concerning the mechanics of proofs and representation of results, the main difference between the models stems from the possible existence of strict bottlenecks in open networks.

As in Section 4, we consider a sequence of open networks with triplets $(\lambda^{0,n}, P, \mu^n)$ and initial queue-length vectors $Q^n(0)$, $n = 1, 2, \ldots$. Here $\lambda^{0,n} \geq 0$, P is substochastic with $\sigma(P) < 1$, $\mu^n > 0$ and $Q^n(0) \geq 0$. In contrast to closed networks, where n parametrizes the population size, now n is merely a rescaling parameter which bears little physical significance. We shall use n in a way that conforms with standard Brownian FCLT's and which differs from the rescaling scheme employed in Theorem 4.1. Such standard rescaling helps clarify the relation between our results and those that already exist [mainly Reiman (1984), Johnson (1983) and Harrison and Williams (1987)].

6.1. For the nth network, the exogenous arrival process is given by $A^n = \{A^n(t) = A^0(\lambda^{0,n}t),\ t \geq 0\}$, the service process by $S^n = \{S^n(t) = S^0(\mu^n t),\ t \geq 0\}$ and the routing sequence R by 2.2.C–2.2.D. The queue lengths Q^n and busy-times B^n are constructed via (2.2)–(2.3). We assume that for some J-dimensional vector c^{λ^0}, c^{μ} and a random vector $\hat{Q}(0) \geq 0$, the following limits exist as $n \to \infty$:

6.1.A $$\sqrt{n}\,(\lambda^{0,n} - \lambda^0) \to c^{\lambda^0},$$

6.1.B $$\sqrt{n}\,(\mu^n - \mu) \to c^{\mu},$$

6.1.C $$\frac{1}{\sqrt{n}} Q^n(0) \to_d \hat{Q}(0).$$

In the formulation of the theorem we use a J-dimensional driftless Brownian motion

6.1.D $$\hat{\xi} = \mathrm{BM}(0, \hat{\Lambda})$$

which starts at $\hat{\xi}(0) = 0$. The covariance matrix $\hat{\Lambda} = [\hat{\Lambda}_{jk}]$ is given by

6.1.E $$\begin{aligned}\hat{\Lambda}_{jk} = {}& \left[\lambda_j^0\left(a_j^2 - 1\right) + \lambda_j + (\lambda_j \wedge \mu_j) b_j^2\right]\delta_{jk} \\ & -(\lambda_j \wedge \mu_j) b_j^2 p_{jk} - (\lambda_k \wedge \mu_k) b_k^2 p_{kj} \\ & -\sum_{l=1}^{J} (\lambda_l \wedge \mu_l) p_{lj} p_{lk}\left[1 - b_l^2\right],\end{aligned}$$

where λ is the inflow capacity vector of the open network (λ^0, P, μ), as determined by (2.1). We maintain the "bar" convention that was introduced for closed networks in Subsection 4.4. The "hat" convention changes, however, because we rescale open networks differently. Indeed, our diffusion limits for open networks arise as time is accelerated by a factor of n while space is aggregated by a factor of $\sqrt{n}$. In accordance with this rescaling, let

$$\begin{aligned}\hat{Q}^n(t) &= \sqrt{n}\left[\bar{Q}^n(t) - (\lambda - \mu)^+ t\right], & \hat{W}^n(t) &= \sqrt{n}\left[\bar{W}^n(t) - (\rho - e)^+ t\right], \\ \hat{D}^n_{j,h}(t) &= \sqrt{n}\,\bar{D}^n_{j,h}(t), & \hat{B}^n(t) &= \sqrt{n}\left[\bar{B}^n(t) - (\rho^n \wedge e)t\right], \\ \hat{S}^n(t) &= \sqrt{n}\left[\bar{S}^n(t) - \mu^n t\right], & \hat{V}^n_j(t) &= \sqrt{n}\left[\bar{V}^n_j(t) - t\right]/\mu^n_j, \\ \hat{A}^n(t) &= \sqrt{n}\left[\bar{A}^n(t) - \lambda^{0,n} t\right], & \hat{R}^n(t) &= \sqrt{n}\left[\bar{R}^n(t) - P't\right].\end{aligned}$$

Here ρ is the traffic intensity vector of the network (λ^0, P, μ). Recalling that the sets α, β and γ stand for nonbottleneck, balanced and strict bottleneck stations, respectively, we now have:

THEOREM 6.1. *Consider the above sequence of open networks. Assume that* 6.1.A–6.1.C *hold and let* $\hat{\xi}$ *be the Brownian motion* 6.1.D. *Then the weak convergence*

$$\left(\hat{Q}^n, \hat{W}^n, \hat{B}^n, \hat{D}^n\right) \to_d \left(\hat{Q}, \hat{W}, \hat{B}, \hat{D}\right) \quad \text{in } t > 0, \tag{6.1}$$

holds as $n \to \infty$. *The limit is described for* $t > 0$ *by*

$$\hat{Q}_\alpha = 0, \tag{6.2}$$

$$\hat{Q}_\beta = X + \left[I - \tilde{P}'_\beta\right]Y, \tag{6.3}$$

$$\begin{aligned} X(t) = X(0) &+ \hat{\xi}_\beta(t) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{\xi}_\alpha(t) \\ &+ \left\{\tilde{c}_\beta^{\lambda^0} + \left[\tilde{P}'_\beta - I\right]c_\beta^\mu + \tilde{P}'_{\gamma\beta}c_\gamma^\mu\right\}t, \end{aligned} \tag{6.4}$$

$$X(0) = \hat{Q}_\beta(0) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{Q}_\alpha(0), \tag{6.5}$$

$$\tilde{c}_\beta^{\lambda^0} = c_\beta^{\lambda^0} + P'_{\alpha\beta}[I - P'_\alpha]^{-1}c_\alpha^{\lambda^0}, \tag{6.6}$$

$$\tilde{P}_{\gamma\beta} = P_{\gamma\beta} + P_{\gamma\alpha}[I - P_\alpha]^{-1}P_{\alpha\beta}, \tag{6.7}$$

$$\tilde{P}_\beta = P_\beta + P_{\beta\alpha}[I - P_\alpha]^{-1}P_{\alpha\beta}, \tag{6.8}$$

$$Y = \Psi_{\tilde{P}_\beta}(X), \tag{6.9}$$

$$\hat{Q}_\gamma = \left[\hat{Q}_\gamma(0) + \hat{\xi}_\gamma\right] + P'_{\alpha\gamma}[I - P'_\alpha]^{-1}\left[\hat{Q}_\alpha(0) + \hat{\xi}_\alpha\right] - \tilde{P}'_{\beta\gamma}Y, \tag{6.10}$$

$$\tilde{P}_{\beta\gamma} = P_{\beta\gamma} + P_{\beta\alpha}[I - P_\alpha]^{-1}P_{\alpha\gamma}, \tag{6.11}$$

$$\hat{W} = \operatorname{diag}(\mu^{-1})\left[\hat{Q} - \operatorname{diag}(\rho - e)^+ c^\mu\right], \tag{6.12}$$

$$\hat{D}_{j,h} = \sum_{k \in \beta} \frac{h_k}{\mu_k}\hat{Q}_k = h\hat{W}, \qquad h_\gamma = 0, \tag{6.13}$$

$$\hat{B}_\alpha = \operatorname{diag}(\mu_\alpha^{-1})[I - P'_\alpha]^{-1}\left[\hat{Q}_\alpha(0) + \hat{\xi}_\alpha - P'_{\beta\alpha}Y + P'_{\beta\alpha}\left[c_\beta^\mu - c_\beta^\lambda\right]^+ t\right], \tag{6.14}$$

$$c_\beta^\lambda = \left\{c_\beta^{\lambda^0} + P'_{\alpha\beta}[I - P'_\alpha]^{-1}c_\alpha^{\lambda^0} + \tilde{P}'_{\gamma\beta}c_\gamma^\mu\right\} + \tilde{P}'_\beta\left(c_\beta^\lambda \wedge c_\beta^\mu\right), \tag{6.15}$$

$$\hat{B}_\beta = -\operatorname{diag}\left(\mu_\beta^{-1}\right)Y, \tag{6.16}$$

$$\hat{B}_\gamma = 0. \tag{6.17}$$

6.2. *Remarks.*

REMARK 1. An obvious modification of Remark 1 that follows Theorem 4.1 applies here as well. The drift vector of X is now

$$\delta = \tilde{c}_\beta^{\lambda^0} + \left[\tilde{P}'_\beta - I\right]c_\beta^\mu + \tilde{P}'_{\gamma\beta}c_\gamma^\mu. \tag{6.18}$$

REMARK 2. Theorem 6.1 again demonstrates that the diffusion limits of queue lengths and workloads vanish at the nonbottlenecks α. The diffusion approximation of the balanced subnetwork β is a $|\beta|$-dimensional open RMB $[\sigma(\tilde{P}_\beta) < 1$ by Part (c) of Lemma 4.3 in CMa]. The diffusion limits for queue lengths at strict bottlenecks require centering. This is because Q_γ^n builds up at a rate of $\lambda_\gamma - \mu_\gamma$, which is also the equilibrium buildup rate of the corresponding fluid approximation [cf. Subsection 3.2]. After both centering and rescaling, Q_γ^n converges to a semimartingale, as apparent from (6.10): its martingale component is a Brownian motion which is associated with α and γ; its bounded-variation component is nonincreasing and is associated with β [see (6.9)].

Remark 3. In analogy to Remark 3 that pertains to Theorem 4.1, the convergence (6.1) fails to hold in $t \geq 0$ when $\hat{Q}_\alpha(0) \neq 0$; Remark 4 there applies after adding to $\hat{Q}_\alpha(0) = 0$ the assumption $\hat{Q}_\gamma(0) = 0$; Remark 8 is valid with $\lambda_\beta^{0,n} = \lambda_\beta^0$ and $\mu_\beta^n = \mu_\beta$.

Remark 4. The constant vector that is substracted from $\hat{Q}$ in (6.12) arises from the particular centering of $\hat{W}^n$ and $\hat{Q}^n$. It would not have appeared if the centering was around $(\rho^n - e)^+$ and $(\lambda^n - \mu^n)^+$, rather than the present $(\rho - e)^+$ and $(\lambda - \mu)^+$, respectively. [Here λ^n and ρ^n are the inflow capacity and traffic intensity vectors of the nth network $(\lambda^{0,n}, P, \mu^n)$.] With this alternative centering, however, the drift (6.18) of X in (6.4) would have been more complicated.

Remark 5. In matrix form, the covariance matrix of $\hat{\xi}$ is

$$\hat{\Lambda} = \Gamma^0[A - I] + \operatorname{diag}(\lambda) + \Gamma B - \Gamma B P - P'B\Gamma - P'[I - B]\Gamma P,$$

where

$$\Gamma = \operatorname{diag}(\lambda \wedge \mu), \qquad B = \operatorname{diag}(b_1^2, \dots, b_J^2),$$
$$\Gamma^0 = \operatorname{diag}(\lambda^0), \qquad A = \operatorname{diag}(a_1^2, \dots, a_J^2).$$

Remark 6. Theorem 6.1 provides light-traffic approximations for open networks without bottlenecks ($\beta = \gamma = \varnothing$). For example, the rescaled busy-time processes $\hat{B}^n(t)$ converge weakly, in this case, to the driftless Brownian motion

$$\hat{B} = \operatorname{diag}(\mu^{-1})\big[\hat{Q}(0) + \hat{\xi}\big],$$

where $\hat{\xi}$ is defined in 6.1.D. (Note that $\lambda_j < \mu_j$ for all j simplifies 6.1.E.)

Remark 7. The equality (6.15) is, in fact, a nonlinear equation with c_β^λ the unknown. Lemma 3.2 in CMa guarantees that (6.15) indeed determines c_β^λ uniquely. In analogy with Remark 9 after Theorem 4.1, (6.15) is the second equality in the following:

Lemma 6.2. *Let* λ^n *and* λ *be the traffic intensities of the networks* $(\lambda^{0,n}, P, \mu^n)$ *and* (λ^0, P, μ), *respectively. If* 6.1.A *and* 6.1.B *hold, then the* *sequence* $\sqrt{n}\,[\lambda^n - \lambda]$ *converges as* $n \to \infty$. *Its limit*, c^λ, *is given by*

$$c_\alpha^\lambda = [I - P'_\alpha]^{-1}\left[c_\alpha^{\lambda^0} + P'_{\beta\alpha}\left(c_\beta^\lambda \wedge c_\beta^\mu\right) + P'_{\gamma\alpha}c_\gamma^\mu\right],$$

$$c_\beta^\lambda = \left\{c_\beta^{\lambda^0} + P'_{\alpha\beta}[I - P'_\alpha]^{-1}c_\alpha^{\lambda^0} + \tilde{P}'_{\gamma\beta}c_\gamma^\mu\right\} + \tilde{P}'_\beta\left(c_\beta^\lambda \wedge c_\beta^\mu\right),$$

$$c_\gamma^\lambda = c_\gamma^{\lambda^0} + P'_{\alpha\gamma}c_\alpha^\lambda + P'_{\beta\gamma}\left(c_\beta^\lambda \wedge c_\beta^\mu\right) + P'_\gamma c_\gamma^\mu.$$

To prove Lemma 6.2, one verifies first that the sequence $\sqrt{n}\,(\lambda^n - \lambda)$ is bounded. Then one proves that its limit points solve the three equalities above. The uniqueness of the solution now establishes the lemma.

6.3. The proof of Theorem 6.1 is carried out, as in Subsection 4.2, under the following assumptions. With probability 1 as $n \to \infty$:

6.3.A $\hat{Q}^n(0) \to \hat{Q}(0)$,

6.3.B $\hat{A}^{0,n}(t) = \sqrt{n}\left[\bar{A}^{0,n}(t) - \lambda^{0,n}t\right] \to \hat{A}^0(t)$ u.o.c.,

6.3.C $\hat{S}^n(t) = \sqrt{n}\left[\bar{S}^n(t) - \mu^n t\right] \to \hat{S}(t)$ u.o.c.,

6.3.D $\hat{R}^n(t) = \sqrt{n}\left[\bar{R}^n(t) - P't\right] \to \hat{R}(t)$ u.o.c.,

6.3.E $\hat{V}_j^n(t) = \sqrt{n}\left[\bar{V}_j^n(t) - t\right]/\mu_j^n \to \hat{V}_j(t)$ u.o.c. for $j = 1, \dots, J$.

Here

$\hat{A}^0 = \mathrm{BM}(0, \Lambda_A)$ with $(\Lambda_A)_{jk} = \delta_{jk}\lambda_j^0 a_j^2$,
$\hat{S}(t) = \mathrm{BM}(0, \Lambda_S)$ with $(\Lambda_S)_{jk} = \delta_{jk}\mu_j b_j^2$,
$\hat{R} = (\hat{R}^1, \dots, \hat{R}^J)$, where $\hat{R}^j = \mathrm{BM}(0, \Lambda_R^j)$ with $\left(\Lambda_R^j\right)_{kl} = p_{jk}\left(\delta_{kl} - p_{jl}\right)$,

$\hat{V}_j(t) = -\hat{S}_j(t/\mu_j)/\mu_j$.

In view of 6.1.A and 6.1.B, the convergences 6.3.B–6.3.D, respectively, imply that

6.3.F $\bar{A}^{0,n}(t) \to \lambda^0 t$ u.o.c.,

6.3.G $\bar{S}^n(t) \to \mu t$ u.o.c.,

6.3.H $\bar{R}^n(t) \to P't$ u.o.c.

We are now ready to prove Theorem 6.1.

6.4. The starting point is the representation 2.7.A, namely

6.4.A $Q^n(t) = X^n(t) + [I - P']Y^n(t)$,

where

$$X^n(t) = Q^n(0) + \theta^n \sqrt{n}\, t + \xi^n(t), \tag{6.19}$$

$$\theta^n = \lambda^{0,n} - [P' - I]\mu^n, \tag{6.20}$$

$$\begin{aligned} \xi_j^n(t) &= \left[A_j^{0,n}(t) - \lambda_j^{0,n} t\right] \\ &\quad + \sum_{k=1}^{J} \left\{R_j^k[S_k^n(B_k^n(t))] - p_{kj}\mu_k^n B_k^n(t)\right\} \\ &\quad - \left[S_j^n(B_j^n(t)) - \mu_j^n B_j^n(t)\right], \qquad j = 1, \dots, J, \end{aligned} \tag{6.21}$$

$$Y_j^n(t) = \mu_j^n\left[t - B_j^n(t)\right], \tag{6.22}$$

$$B_j^n(t) = \int_0^t 1\left[Q_j^n(u) > 0\right] du. \tag{6.23}$$

Similarly to the proof of Theorem 4.1, one starts with proving the analogs of Propositions 4.3 and 4.4. The latter is rather immediate. Indeed, the convergence 6.3.A implies that $\overline{Q}^n(0) \to 0$ as $n \to \infty$. Thus, the analog of Proposition 4.4 follows from CMa (Theorem 5.1 and Remark 2 succeeding Theorem 5.2), namely

$$\overline{B}^n(t) \to (\rho \wedge e)t \quad \text{u.o.c. as } n \to \infty. \tag{6.4.B}$$

We next establish the analog of Proposition 4.3:

$$\hat{Q}_\alpha^n\left(t + \frac{\tau}{\sqrt{n}}\right) \to 0 \quad \text{u.o.c. as } n \to \infty, \tag{6.4.C}$$

for some $\tau \geq 0$ which will be specified below. To this end recall the derivation of (4.14)–(4.15) from 4.5.C–4.5.D. Then deduce from 6.3.B–6.3.D the existence of the following three limits as $n \to \infty$:

$$\frac{1}{\sqrt{n}} A^{0,n}(\sqrt{n}\, t) \to \lambda^0 t \quad \text{u.o.c.}, \tag{6.24}$$

$$\frac{1}{\sqrt{n}} S^n(\sqrt{n}\, t) \to \mu t \quad \text{u.o.c.}, \tag{6.25}$$

$$\frac{1}{\sqrt{n}} R\left(\left[\sqrt{n}\, t\right]\right) \to P't \quad \text{u.o.c.} \tag{6.26}$$

Now 6.1.A–6.1.B, 6.3.A and (6.24)–(6.26) provide all the prerequisites for the fluid approximations (Theorems 5.1 and 5.2 both in CMa) of the processes $(1/\sqrt{n})Q^n(\sqrt{n}\, t)$, $(1/\sqrt{n})B^n(\sqrt{n}\, t)$ and $(1/\sqrt{n})Y^n(\sqrt{n}\, t)$. Replacing n with $\sqrt{n}$ in the argument that leads to 5.1.A yields 6.4.C, with $\tau \geq 0$ being the equilibrium time (Theorem 5.2 of CMa) for the process $(1/\sqrt{n})Q^n(\sqrt{n}\, t)$.

6.5. Consider the sequence $\hat{Y}^n$ defined by

$$\hat{Y}^n(t) = \frac{1}{\sqrt{n}} Y^n(nt).$$

The present section is devoted to proving that:

6.5.A Both sequences $\{\hat{Y}_\beta^n,\ n \geq 0\}$ and $\{\hat{Y}_\gamma^n,\ n \geq 0\}$ are equicontinuous and uniformly bounded on any compact subset of $[0, \infty)$.

Introduce $\vec{X}^n(t)$, $\vec{Y}^n(t)$ and $\vec{Q}^n(t)$ by replacing n in (5.1)–(5.3) with $\sqrt{n}$. For example, $\vec{Q}^n(t) = (1/\sqrt{n})Q^n(n(t + \tau/\sqrt{n}))$. Let β' be the union of β and γ. As in Subsection 5.3, write (5.15)–(5.18) with β' substituting for β and $\sqrt{n}\,\tilde{\theta}_{\beta'}^n$ for $n\tilde{\theta}_\beta^n$. It follows from 6.6.C–6.6.D, which will be verified independently later, that $\sqrt{n}\,\tilde{\theta}_{\beta'}^n$ is bounded below, say by r. Consequently $\tilde{X}_{\beta'}^n$ is bounded below by

$$\tilde{\chi}_{\beta'}^n(t) = \tilde{Q}_{\beta'}^n(0) + rt + \tilde{\xi}_{\beta'}^n(t).$$

From the characterization 3.3.F of the regulator as a least element it follows that

$$0 \leq \vec{Y}_{\beta'}^n = \Psi_{\tilde{P}_{\beta'}}\left(\tilde{X}_{\beta'}^n\right) \leq \Psi_{\tilde{P}_{\beta'}}\left(\tilde{\chi}_{\beta'}^n\right). \tag{6.27}$$

Now prove, as in Subsection 5.3, that $\tilde{\chi}_{\beta'}^n$ converges u.o.c. to a process in the domain of $\Psi_{\tilde{P}_{\beta'}}$. By 3.3.D, the sequence $\Psi_{\tilde{P}_{\beta'}}(\tilde{\chi}_{\beta'}^n)$ converges u.o.c. to a limit which is nondecreasing, continuous and vanishes at $t = 0$. The monotonicity of $\hat{Y}^n(\cdot)$, together with (6.27), implies that $\hat{Y}_{\beta'}^n$ is equicontinuous at $t = 0$. To treat an arbitrary $t_0 > 0$, apply the Markovian property 3.5.A–3.5.C to establish an upper bound similar to (6.27) (which is equicontinuous) for $\vec{Y}^n(t) - \vec{Y}^n(t_0)$, $t \geq t_0$. We conclude that $\hat{Y}^n$ is equicontinuous, thus verifying 6.5.A.

6.6. In the present subsection we prove that

6.6.A $$\hat{Y}_\gamma^n \to 0 \quad \text{u.o.c.},$$

starting with some preparatory results. As in verifying (5.22), one shows that $\hat{\xi}_j^n$ converges u.o.c. to

$$\begin{aligned}\hat{\xi}_j(t) = {}& \hat{A}_j^0(t) + \sum_{k=1}^{J} \hat{R}_j^k((\lambda_k \wedge \mu_k)t) \\ & + \sum_{k=1}^{J} \hat{S}_k((\rho_k \wedge 1)t)p_{kj} - \hat{S}_j((\rho_j \wedge 1)t),\end{aligned} \tag{6.28}$$

for $j = 1, \dots, J$. Now write the hat version of 6.4.A in blocks α, β and γ:

$$\hat{Q}_\alpha^n = \hat{X}_\alpha^n - P'_{\beta\alpha}\hat{Y}_\beta^n - P'_{\gamma\alpha}\hat{Y}_\gamma^n + [I - P'_\alpha]\hat{Y}_\alpha^n, \tag{6.29}$$

$$\hat{Q}_\beta^n = \hat{X}_\beta^n - P'_{\alpha\beta}\hat{Y}_\alpha^n - P'_{\gamma\beta}\hat{Y}_\gamma^n + \left[I - P'_\beta\right]\hat{Y}_\beta^n, \tag{6.30}$$

$$\hat{Q}_\gamma^n + \left[\lambda_\gamma - \mu_\gamma\right]\sqrt{n}\,t = \hat{X}_\gamma^n - P'_{\alpha\gamma}\hat{Y}_\alpha^n - P'_{\beta\gamma}\hat{Y}_\beta^n + \left[I - P'_\gamma\right]\hat{Y}_\gamma^n. \tag{6.31}$$

Since $\rho(P) < 1$, the inverse of $[I - P'_\alpha]$ exists [Corollary 2.1.6 and Lemma 6.2.1 in Berman and Plemmons (1979)]. Solving for $\hat{Y}_\alpha^n$ in (6.29) and substituting the outcome into (6.30) and (6.31) yields

$$\hat{Q}_\beta^n = \tilde{X}_\beta^n - \tilde{P}'_{\gamma\beta}\hat{Y}_\gamma^n + \left[I - \tilde{P}'_\beta\right]\hat{Y}_\beta^n, \tag{6.32}$$

where

$$\begin{aligned}
(6.33)\qquad \tilde{X}^n_\beta(t) &= \tilde{Q}^n_\beta(0) + \tilde{\xi}^n_\beta(t) + \tilde{\theta}^n_\beta\sqrt{n}\,t - P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{Q}^n_\alpha(t),\\
\tilde{Q}^n_\beta(0) &= \hat{Q}^n_\beta(0) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{Q}^n_\alpha(0),\\
\tilde{\xi}^n_\beta(t) &= \hat{\xi}^n_\beta(t) + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\hat{\xi}^n_\alpha(t),\\
\tilde{\theta}^n_\beta &= \theta^n_\beta + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\theta^n_\alpha,
\end{aligned}$$

$\tilde{P}_{\gamma\beta}$ and $\tilde{P}_\beta$ are defined in (6.7) and (6.8), respectively, and

$$(6.34)\qquad \hat{Q}^n_\gamma + \left[\lambda_\gamma - \mu_\gamma\right]\sqrt{n}\,t = \tilde{X}^n_\gamma - \tilde{P}'_{\beta\gamma}\hat{Y}^n_\beta + \left[I - \tilde{P}'_\gamma\right]\hat{Y}^n_\gamma,$$

where

$$\begin{aligned}
(6.35)\qquad \tilde{X}^n_\gamma(t) &= \tilde{Q}^n_\gamma(0) + \tilde{\xi}^n_\gamma(t) + \tilde{\theta}^n_\gamma\sqrt{n}\,t - P'_{\alpha\gamma}[I - P'_\alpha]^{-1}\hat{Q}^n_\alpha(t),\\
\tilde{Q}^n_\gamma(0) &= \hat{Q}^n_\gamma(0) + P'_{\alpha\gamma}[I - P'_\alpha]^{-1}\hat{Q}^n_\alpha(0),\\
\tilde{\xi}^n_\gamma(t) &= \hat{\xi}^n_\gamma(t) + P'_{\alpha\gamma}[I - P'_\alpha]^{-1}\hat{\xi}^n_\alpha(t),\\
\tilde{\theta}^n_\gamma &= \theta^n_\gamma + P'_{\alpha\gamma}[I - P'_\alpha]^{-1}\theta^n_\alpha,\\
\tilde{P}_{\beta\gamma} &= P_{\beta\gamma} + P_{\beta\alpha}[I - P_\alpha]^{-1}P_{\alpha\gamma},\\
\tilde{P}_\gamma &= P_\gamma + P_{\gamma\alpha}[I - P_\alpha]^{-1}P_{\alpha\gamma}.
\end{aligned}$$

From 6.1.A and 6.1.B it follows that $\sqrt{n}[\theta^n - \theta]$ converges to $c^{\lambda^0} + [P' - I]c^\mu$, where $\theta = \lambda^0 + [P' - I]\mu$. Let

$$\tilde{\theta}_\beta = \theta_\beta + P'_{\alpha\beta}[I - P'_\alpha]^{-1}\theta_\alpha \quad\text{and}\quad \tilde{\theta}_\gamma = \theta_\gamma + P'_{\alpha\gamma}[I - P'_\alpha]^{-1}\theta_\alpha.$$

Writing $\theta = \lambda^0 + [P' - I]\mu$ and the traffic equation (2.1) in blocks α, β and γ, one can show that $\theta_\alpha = [I - P'_\alpha](\lambda_\alpha - \mu_\alpha)$, $\tilde{\theta}_\beta = 0$ and $\tilde{\theta}_\gamma = \lambda_\gamma - \mu_\gamma$. Therefore, as $n \to \infty$,

6.6.B $\sqrt{n}\,[\theta^n_\alpha - \theta_\alpha] \to \tilde{c}^{\lambda^0}_\alpha$,

where $\theta_\alpha = [I - P'_\alpha](\lambda_\alpha - \mu_\alpha)$

and $\tilde{c}^{\lambda^0}_\alpha = c^{\lambda^0}_\alpha + [P'_\alpha - I]c^\mu_\alpha + P'_{\beta\alpha}c^\mu_\beta + P'_{\gamma\alpha}c^\mu_\gamma$;

6.6.C $\sqrt{n}\left[\tilde{\theta}^n_\beta - \tilde{\theta}_\beta\right] \to \tilde{c}^{\lambda^0}_\beta + \left[\tilde{P}'_\beta - I\right]c^\mu_\beta + \tilde{P}'_{\gamma\beta}c^\mu_\gamma$,

where $\tilde{\theta}_\beta = 0$ and $\tilde{c}^{\lambda^0}_\beta = c^{\lambda^0}_\beta + P'_{\alpha\beta}[I - P'_\alpha]^{-1}c^{\lambda^0}_\alpha$;

6.6.D $\sqrt{n}\left[\tilde{\theta}^n_\gamma - \tilde{\theta}_\gamma\right] \to \tilde{c}^{\lambda^0}_\gamma + \left[\tilde{P}'_\gamma - I\right]c^\mu_\gamma + \tilde{P}'_{\beta\gamma}c^\mu_\beta$,

where $\tilde{\theta}_\gamma = \lambda_\gamma - \mu_\gamma > 0$ and $\tilde{c}^{\lambda^0}_\gamma = c^{\lambda^0}_\gamma + P'_{\alpha\gamma}[I - P'_\alpha]^{-1}c^{\lambda^0}_\alpha$.

Proof of 6.6.A. Let $\chi^n_\gamma = \tilde{X}^n_\gamma - \tilde{P}'_{\beta\gamma}\hat{Y}^n_\beta$, with $\tilde{X}^n_\gamma$ from (6.35). By 6.5.A and 6.6.D,

$$(6.36)\qquad \chi^n_\gamma \to +\infty \quad \text{u.o.c. in } t > 0 \text{ as } n \to \infty.$$

Again by 6.5.A, now combined with the theorem of Arzela–Ascoli, any subsequence of $\hat{Y}^n_\gamma$ has a further subsequence that converges u.o.c. to a nondecreasing continuous function, say $\hat{Y}_\gamma$. Denote the index of this subsequence by n_k,

$k = 1, 2, \ldots$. If $\hat{Y}_\gamma \not\equiv 0$, there must be an $N \geq 1$ and $t_1 > t_0 > 0$ such that

$$\hat{Y}_j^{n_k}(t_1) > \hat{Y}_j^{n_k}(t_0) \quad \text{for all } k \geq N,$$

for some $j \in \gamma$. By property 3.3.C,

$$\frac{1}{\sqrt{n_k}} Q_j^{n_k}(n_k s_k) = 0 \quad \text{for all } k \geq N, \tag{6.37}$$

where $s_k \in [t_0, t_1]$. On the other hand, we see from (6.34) that 6.5.A and (6.36) imply

$$\frac{1}{\sqrt{n}} Q_j^n(nt) \to +\infty, \quad \text{uniformly in } t \in [t_0, t_1],$$

which contradicts (6.37). Therefore, $\hat{Y}_\gamma \equiv 0$, proving 6.6.A. □

6.7. *The convergence of $\hat{Q}^n$ and $\hat{B}^n$.* First one proves the convergence of $\hat{Q}_\beta^n$ and $\hat{Y}_\beta^n$ by an argument similar to the one in Subsection 5.4. One must add 6.6.A because of the additional term $\hat{Y}_\gamma^n$ in (6.32). In view of (6.34), the convergences of $\hat{\xi}_\gamma^n$ and $\hat{Y}_\beta^n$, together with 6.6.A and 6.6.D, imply the convergence of $\hat{Q}_\gamma^n$, thus establishing the $\hat{Q}^n$ part.

The proof of convergence for $\hat{B}_\gamma^n$ is given by 6.6.A. The convergence of $\hat{B}_\beta^n$ follows from $\hat{B}_\beta^n = -\operatorname{diag}(\mu_\beta^{-1})\hat{Y}_\beta^n$ and the convergence of $\hat{Y}_\beta^n$. Finally, one is left with verifying the convergence of $\hat{B}_\alpha^n$. In view of 6.6.B, rewrite (6.29) as

$$\begin{aligned}
&\left[\hat{Y}_\alpha^n(t) - \sqrt{n}(\mu_\alpha^n - \lambda_\alpha^n)t\right] \\
&\quad = -[I - P'_\alpha]^{-1}\Big\{\hat{Q}_\alpha^n(0) - \hat{Q}_\alpha^n(t) + \hat{\xi}_\alpha^n(t) \\
&\qquad\qquad - P'_{\beta\alpha}\hat{Y}_\beta^n - P'_{\gamma\alpha}\hat{Y}_\gamma^n + \sqrt{n}(\theta_\alpha^n - \theta_\alpha)t \\
&\qquad\qquad + [I - P'_\alpha]\sqrt{n}\left[(\mu_\alpha^n - \mu_\alpha) - (\lambda_\alpha^n - \lambda_\alpha)\right]t\Big\}.
\end{aligned} \tag{6.38}$$

By 6.4.C, (6.28), 6.6.A, 6.6.B, Lemma 6.2 and the convergence of $\hat{Y}_\beta^n$ to Y_β, we prove via (6.38) that as $n \to \infty$,

$$\begin{aligned}
&\hat{Y}_\alpha^n - \sqrt{n}(\mu_\alpha^n - \lambda_\alpha^n)t \\
&\quad \to -[I - P'_\alpha]^{-1}\left[\hat{Q}_\alpha(0) + \hat{\xi}_\alpha - P'_{\beta\alpha}Y + P'_{\beta\alpha}\left[c_\beta^\mu - c_\beta^\lambda\right]^+ t\right] \quad \text{u.o.c.},
\end{aligned}$$

and the proof is completed in view of the relation

$$\hat{B}_\alpha^n(t) = -\operatorname{diag}(\mu_\alpha^{-1})\left[\hat{Y}_\alpha^n - \sqrt{n}(\mu_\alpha^n - \lambda_\alpha^n)t\right].$$

6.8. *The convergence of $\hat{W}^n$ and $\hat{D}^n$.* The proofs are essentially the same as those in Subsections 5.5 and 5.6. Major differences are the different time and space rescaling, and the fact that at the right-hand side of relation (5.17), a term $(\rho_j - 1)^+\sqrt{n}[\mu_j - \mu_j^n]/\mu_j^n$, which converges to $-(\rho_j - 1)^+ c_j^\mu/\mu_j$ as $n \to \infty$, must be added.

7. Networks with priorities. The results obtained so far can be extended to accommodate some networks with a nonhomogeneous customer population. (Other such networks, the analysis of which is beyond the present scope, are described in Subsection 8.9.) We shall now derive diffusion approximations for what we call *prioritized* closed networks. These are closed networks in which two *types* of customers circulate: type h for high priority and type l for low priority. Each type has its own service and routing characteristics and they interact solely through contention over service in the following manner. An l-customer gets served at a station only when there are no h-customers present there. Suppose that an h-customer arrives at a station amidst service of an l-customer. Then the service is interrupted, the server immediately attends to the h-customer, the l-customer is forced to return to the queue and he is back up for service precisely when the station is again empty of h-customers. Finally, an interrupted service of an l-customer resumes from the point of interruption, rather than starting afresh. Such a state of affairs is often summarized by saying that h-customers enjoy a preemptive-resume priority over l-customers.

In Theorem 7.1 we present diffusion limits for the queue lengths and busy times of both types of customers. As before, fluid approximations are a prerequisite for proving diffusion limits and these are described in Remark 2 following Theorem 7.1. The diffusion and fluid approximations both arise as the total population size increases indefinitely, while maintaining the number of customers from each type comparable. For simplicity, it is assumed that the service times and routing indicators do *not* vary with the population size. We chose to analyze a closed network because open networks were already partially covered by Johnson (1983) (see Subsection 8.4 for more details). Restricting the attention to two types of customers facilitates the presentation considerably, while still providing all the ideas and machinery required for the analysis of three types or more.

7.1. Entities associated with h- and l-customers will be appended with an h and an l, respectively. For example, $Q_j^h(t)$ represents the queue length of h-customers at station j at time t and $B_j^l(t)$ stands for the cumulative time allocated by server j to serve l-customers during the time interval $[0, t]$. The triplets associated with the types are denoted by $(0, P(h), \mu^h)$ and $(0, P(l), \mu^l)$, where $P(h)$ and $P(l)$ are assumed irreducible.

The dynamics of h-customers, unaffected by l-customers, are identical to those in Section 4. In particular [cf. (2.2) and (2.3)], the queue length and busy-time process jointly satisfy, for $j = 1, \dots, J$,

$$\text{7.1.A} \quad Q_j^h(t) = Q_j^h(0) + \sum_{k=1}^{J} R_j^{h,k}\{S_k^h[B_k^h(t)]\} - S_j^h[B_j^h(t)], \qquad t \geq 0,$$

$$\text{7.1.B} \quad B_j^h(t) = \int_0^t 1[Q_j^h(u) > 0]\, du, \qquad t \geq 0.$$

Then for l-customers and $j = 1, \dots, J$ we have

$$\text{7.1.C} \quad Q_j^l(t) = Q_j^l(0) + \sum_{k=1}^{J} R_j^{l,k}\{S_k^l[B_k^l(t)]\} - S_j^l[B_j^l(t)], \qquad t \geq 0,$$

$$\text{7.1.D} \quad B_j^l(t) = \int_0^t 1[Q_j^h(u) = 0, Q_j^l(u) > 0]\, du, \qquad t \geq 0.$$

[One can prove that, given $Q^h(t)$, the processes $Q^l(t)$ and $B^l(t)$ indeed exist and are uniquely determined by 7.1.C–7.1.D.] For stating the limit theorem, it is convenient to introduce the total queue length and busy-time processes

$$Q(t) = Q^h(t) + Q^l(t), \qquad B(t) = B^h(t) + B^l(t).$$

Note that Q and B are related by

$$B_j(t) = \int_0^t 1[Q_j(u) > 0]\,du, \qquad t \geq 0,$$

for $j = 1, \dots, J$.

7.2. We shall analyze a sequence of prioritized closed networks indexed by their total population size n. As usual, entities associated with the nth network are appended with a superscript n. Thus

$$e'Q^n(0) = e'Q^{h,n}(0) + e'Q^{l,n}(0) = n.$$

We further suppose that the following limits exist as $n \to \infty$:

7.2.A $$\frac{1}{n}Q^{h,n}(0) \to_d \hat{Q}^h(0) \quad \text{with} \quad e'\hat{Q}^h(0) > 0,$$

7.2.B $$\frac{1}{n}Q^{l,n}(0) \to_d \hat{Q}^l(0) \quad \text{with} \quad e'\hat{Q}^l(0) > 0.$$

(To avoid complications of no significance, the latter two inequalities are taken to be satisfied with probability 1.) In the formulation of the theorem we use a J-dimensional driftless Brownian motion

7.2.C $$\hat{\xi}^h = \mathrm{BM}(0, \hat{\Lambda})$$

which starts at $\hat{\xi}^h(0) = 0$. The covariance matrix $\hat{\Lambda} = [\hat{\Lambda}_{jk}]$ is given by

7.2.D $$\hat{\Lambda}_{jk} = \lambda_j^h \delta_{jk}\left[1 + \left(b_j^h\right)^2\right] - \lambda_j^h\left(b_j^h\right)^2 p_{jk}(h) - \lambda_k\left(b_k^h\right)^2 p_{kj}(h)$$
$$- \sum_{l=1}^{j} \lambda_l^h p_{lj}(h) p_{lk}(h)\left[1 - \left(b_l^h\right)^2\right],$$

where $\lambda^h = (\lambda_1^h, \dots, \lambda_J^h)'$ is the inflow capacity vector of the closed network $(0, P(h), \mu^h)$, as defined via (2.1). The limit theorem is a FCLT jointly for the sequences

$$\hat{Q}^{h,n}(t) = \frac{1}{n}Q^{h,n}(n^2 t), \qquad \hat{B}^{h,n}(t) = \frac{1}{n}\left[B^{h,n}(n^2 t) - \rho^h nt\right],$$
$$\hat{Q}^{l,n}(t) = \frac{1}{n}Q^{l,n}(n^2 t), \qquad \hat{B}^{l,n}(t) = \frac{1}{n}B^{l,n}(n^2 t),$$
$$\hat{Q}^{n}(t) = \frac{1}{n}Q^{n}(n^2 t), \qquad \hat{B}^{n}(t) = \frac{1}{n}\left[B^{n}(n^2 t) - \rho^h nt\right].$$

Here ρ^h is the traffic intensity of the network $(0, P(h), \mu^h)$. Denote by α and

β, respectively, the set of nonbottleneck and bottleneck stations of the network $(0, P(h), \mu^h)$. Then let

$$\Delta^h = \operatorname{diag}(\mu^h), \qquad \Delta^l = \operatorname{diag}(\mu^l),$$

to ease the notational burden in Theorem 7.1.

THEOREM 7.1. *Consider a sequence of prioritized closed networks indexed by their total population size n. Assume that the service times and routing indicators do not vary with n and that 7.2.A–7.2.B hold. Let $\hat{\xi}^h$ be the Brownian motion 7.2.C. Then the weak convergence*

$$\left(\hat{Q}^{h,n}, \hat{Q}^{l,n}, \hat{Q}^n, \hat{B}^{h,n}, \hat{B}^{l,n}, \hat{B}^n\right) \to_d \left(\hat{Q}^h, \hat{Q}^l, \hat{Q}, \hat{B}^h, \hat{B}^l, \hat{B}\right) \quad \textit{in } t > 0, \tag{7.1}$$

holds as $n \to \infty$. The limit is described for $t > 0$ by

$$\hat{Q}^h_\alpha = 0, \tag{7.2}$$

$$\hat{Q}^h_\beta = X^h + \left[I - \tilde{P}'(h)\right]Y^h, \tag{7.3}$$

$$X^h(t) = X^h(0) + \hat{\xi}^h_\beta(t) + P'_{\alpha\beta}(h)[I - P'_\alpha(h)]^{-1}\hat{\xi}^h_\alpha(t), \tag{7.4}$$

$$X^h(0) = \hat{Q}^h_\beta(0) + P'_{\alpha\beta}(h)[I - P'_\alpha(h)]^{-1}\hat{Q}^h_\alpha(0), \tag{7.5}$$

$$\tilde{P}(h) = P_\beta(h) + P_{\beta\alpha}(h)[I - P_\alpha(h)]^{-1}P_{\alpha\beta}(h), \tag{7.6}$$

$$Y^h = \Psi_{\tilde{P}(h)}(X^h), \tag{7.7}$$

$$\hat{B}^h_\alpha = (\Delta^h_\alpha)^{-1}[I - P'_\alpha(h)]^{-1}\left[\hat{Q}^h_\alpha(0) + \hat{\xi}^h_\alpha - P'_{\beta\alpha}(h)Y\right], \tag{7.8}$$

$$\hat{B}^h_\beta = -\left(\Delta^h_\beta\right)^{-1}Y^h, \tag{7.9}$$

$$\hat{Q}_\alpha = 0, \tag{7.10}$$

$$\hat{Q}_\beta = X + \left[I - \tilde{P}'(l)\right]Y, \tag{7.11}$$

$$X(t) = X(0) + \hat{\xi}_\beta(t) + P'_{\alpha\beta}(l)[I - P'_\alpha(l)]^{-1}\hat{\xi}_\alpha(t), \tag{7.12}$$

$$X(0) = \hat{Q}_\beta(0) + P'_{\alpha\beta}(l)[I - P'_\alpha(l)]^{-1}\hat{Q}_\alpha(0), \tag{7.13}$$

$$\hat{\xi}(t) = \hat{\xi}^h(t) + \{[I - P'(l)]\Delta^l - [I - P'(h)]\Delta^h\}\hat{B}^h(t), \tag{7.14}$$

$$\tilde{P}(l) = P_\beta(l) + P_{\beta\alpha}(l)[I - P_\alpha(l)]^{-1}P_{\alpha\beta}(l), \tag{7.15}$$

$$Y = \Psi_{\tilde{P}(l)}(X), \tag{7.16}$$

$$\hat{B}_\alpha = (\Delta^l_\alpha)^{-1}[I - P'_\alpha(l)]^{-1}\left[\hat{Q}_\alpha(0) + \hat{\xi}_\alpha - P'_{\beta\alpha}(l)Y\right], \tag{7.17}$$

$$\hat{B}_\beta = -\left(\Delta^l_\beta\right)^{-1}Y, \tag{7.18}$$

$$\hat{Q}^l = \hat{Q} - \hat{Q}^h, \tag{7.19}$$

$$\hat{B}^l = \hat{B} - \hat{B}^h. \tag{7.20}$$

7.3. *Remarks.*

REMARK 1. As before, (7.2)–(7.20) are equalities in distribution. The convergence of $\hat{Q}^{n,h}$ and $\hat{B}^{h,n}$ and the relations (7.2)–(7.7) are all consequences of Theorem 4.1. Also, the remarks in Subsection 4.3 all apply to the performance measures of h-customers.

REMARK 2. The bottlenecks β of a prioritized closed network, hence the nonbottlenecks α as well, are determined by the triplet of the h-customers. A deeper support for this remark is provided by the fluid approximation, derived in Subsection 7.5 as an intermediate step along the proof of Theorem 7.1. In the fluid model, each of the J buffers is capable of holding simultaneously, but separately, two types of fluids: h-fluid and l-fluid. The circulation of h-fluid conforms to the rules described in Subsections 3.1–3.3; l-fluid is released from a buffer, according to the same rules, only when h-fluid is not present there. In a finite time, the fluid network reaches the following equilibrium: buffers in α are empty; the h-fluid circulates within β at constant rates λ^h; finally, the l-fluid is motionless because it is never released from a buffer in β once it gets there [see (7.38)]. In a diffusion time scale, equilibrium is reached instantaneously, which explains (7.10).

REMARK 3. A detailed alternative representation of (7.19)–(7.20) is given on $t > 0$ by

$$\hat{Q}^l_\alpha = 0,$$

$$\hat{Q}^l_\beta = \hat{Q}^l_\beta(0) + P'_{\alpha\beta}(l)[I - P'_\alpha]^{-1}\hat{Q}^l_\alpha(0) - \left[I - \tilde{P}'(l)\right]\Delta^l_\beta\hat{B}^l_\beta,$$

$$\hat{B}^l_\alpha = \left(D^l_\alpha\right)^{-1}[I - P'_\alpha(l)]^{-1}\left[\hat{Q}^l_\alpha(0) + P'_{\beta\alpha}(l)\Delta^l_\beta\hat{B}^l_\beta\right],$$

$$\hat{B}^l_\beta = \left(\Delta^h_\beta\right)^{-1}Y^h - \left(\Delta^l_\beta\right)^{-1}Y.$$

7.4. To simplify the proof for Theorem 7.1, let us assume that

7.4.A $$\hat{Q}^h_\alpha(0) = 0 \quad \text{and} \quad \hat{Q}^l_\alpha(0) = 0,$$

in which case (7.1)–(7.20) actually hold in $t \geq 0$. The bar and hat conventions from Section 4 are retained here. In addition, a superscript h or l will always be appended to a process or a property associated with the corresponding customer type. Skorohod's representation is assumed to have been applied. Thus, the primitives associated with the customer types are all defined on a common probability space, there is independence between the types, 4.5.B–4.5.D [hence (4.14) and (4.15)] hold with probability 1 for both customer types and so does 7.2.A–7.2.B. For example, it is assumed that 4.5.C^h holds, meaning that for almost all sample paths, $\hat{S}^{h,n}(t) = \bar{S}^{h,n}(nt) - \mu^h nt \to \hat{S}^h(t)$ u.o.c. as $n \to \infty$.

In preparation for the proof let us center 7.1.A, as in Subsection 2.7, to get

7.4.B $$Q^{h,n} = X^{h,n} + [I - P'(h)]Y^{h,n},$$

where

$$X^{h,n}(t) = Q^{h,n}(0) + \theta^h t + \xi^{h,n}(t), \tag{7.21}$$

$$\theta^h = [P'(h) - I]\mu^h, \tag{7.22}$$

$$\begin{aligned} \xi_j^{h,n}(t) = {} & \sum_{k=1}^{J} \{R_j^h[S_k^h(B_k^{h,n}(t))] - p_{kj}(h)S_k^h(B_k^{h,n}(t))\} \\ & + \sum_{k=1}^{J} p_{kj}(h)[S_k^h(B_k^{h,n}(t)) - \mu_k^h B_k^{h,n}(t)] \\ & - [S_j^h(B_j^{h,n}(t)) - \mu_j^h B_j^{h,n}(t)], \end{aligned} \tag{7.23}$$

$$Y_j^{h,n}(t) = \mu_j^h[t - B_j^{h,n}(t)]. \tag{7.24}$$

The processes $Q^{h,n}$, $X^{h,n}$ and $Y^{h,n}$ uniquely satisfy 2.7.A–2.7.C, playing the roles of Q, X and Y there. From 7.1.A–7.1.D, the total queue-length process can be represented as

7.4.C $$Q^n = X^n + [I - P'(l)]Y^n,$$

where

$$X^n(t) = Q^n(0) + \theta t + \xi^n(t), \tag{7.25}$$

$$\theta = [P'(l) - I]\mu, \tag{7.26}$$

$$\mu = \Delta^l(e - \rho^h), \tag{7.27}$$

$$\begin{aligned} \xi^n(t) = {} & \xi^{h,n}(t) + \xi^{l,n}(t) \\ & + \{[I - P'(h)]\Delta^h - [I - P'(l)]\Delta^l\}[\rho^h - B^{h,n}(t)], \end{aligned} \tag{7.28}$$

$$\begin{aligned} \xi_j^{l,n}(t) = {} & \sum_{k=1}^{J} \{R_j^l[S_k^l(B_k^{l,n}(t))] - p_{kj}(l)S_k^h(B_k^{l,n}(t))\} \\ & + \sum_{k=1}^{J} p_{kj}(l)[S_k^l(B_k^{l,n}(t)) - \mu_k^l B_k^{l,n}(t)] \\ & - [S_j^l(B_j^{l,n}(t)) - \mu_j^l B_j^{l,n}(t)], \end{aligned} \tag{7.29}$$

$$Y_j^n(t) = \mu_j^l[t - B_j^n(t)]. \tag{7.30}$$

As above, Q^n, X^n and Y^n uniquely satisfy 2.7.A–2.7.C. Note that μ, defined in (7.27), represents the service capacity available to l-customers after discounting the capacity allocated to h-customers. The bar and hat analogs of 7.4.B and 7.4.C are omitted for the sake of brevity.

7.5. We now formally describe the fluid model which approximates the prioritized network. One should remark that because of 7.4.A, equilibrium for this model starts immediately at time $t = 0$. In view of Theorem 7.1 in CMa, 4.5.B^h, (4.14)^h, (4.15)^h, 7.4.A and the fact that l-customers do not affect the flow of h-customers, we have

$$\text{7.5.A} \qquad \left(\overline{Q}^{h,n}, \overline{B}^{h,n}\right) \to \left(\overline{Q}^h, \overline{B}^h\right) \quad \text{u.o.c. as } n \to \infty,$$

where

$$(7.31) \qquad \overline{Q}^h(t) = \overline{Q}^h(0) \quad \text{and} \quad \overline{B}^h(t) = \rho^h t \quad \text{for } t \geq 0.$$

Applying 7.5.A, 4.5.B^l, (4.14)^l and (4.15)^l to the representation (7.25) yields

$$(7.32) \qquad \overline{X}^n \to \overline{X} \quad \text{u.o.c. as } n \to \infty,$$

with

$$(7.33) \qquad \begin{aligned} \overline{X}(t) &= \overline{Q}(0) + \theta t = \overline{Q}(0) + [P'(l) - I]\mu t \\ &= \overline{Q}(0) + [P'(l) - I]\Delta^l(e - \rho^h)t. \end{aligned}$$

If μ in (7.33) was a positive vector, then 7.4.C could have been analyzed as arising from the irreducible closed network $(0, P(l), \mu)$. However, $\mu_\beta = 0$ since $\rho^h_\beta = e$, which is circumvented by writing 7.4.C in α and β blocks as

$$(7.34) \qquad Q^n_\alpha = X^n_\alpha - P'_{\beta\alpha}(l)Y^n_\beta + \left[I - P'_\alpha(l)\right]Y^n_\alpha,$$

$$(7.35) \qquad Q^n_\beta = X^n_\beta - P'_{\alpha\beta}(l)Y^n_\alpha + \left[I - P'_\beta(l)\right]Y^n_\beta.$$

From (7.31) and the inequality $\overline{Y}^n_\beta(t) = \Delta^l_\beta[et - \overline{B}^n_\beta(t)] \leq \Delta^l_\beta[et - \overline{B}^{h,n}_\beta(t)]$, it now follows that

$$(7.36) \qquad \overline{Y}^n_\beta \to 0.$$

Corollary 3.29 on page 15 in Berman and Plemmons (1979) guarantees that $\sigma[P_\alpha(l)] < 1$. Since $\mu_\beta = 0$, (7.33) and (7.36) suggest that the representation (7.34) be treated as arising from the open network $(0, P_\alpha(l), \mu_\alpha)$. (Such a network has no exogenous input, hence all its traffic intensities vanish.) Indeed, applying Theorems 5.1 and 5.2 in CMa, while taking into consideration 7.4.A, results in

$$(7.37) \qquad \left(\overline{Q}^n_\alpha(t), \overline{Y}^n_\alpha(t)\right) \to (0, \mu_\alpha t) \quad \text{u.o.c. as } n \to \infty.$$

(The long-run fractions of busy time also vanish.) Using 7.5.A, (7.30) and (7.37), one can calculate now all the performance measures associated with stations in α. For the β part, substitute (7.32), (7.33), (7.36) and (7.37) into (7.35) to conclude that

$$\overline{Q}^n_\beta(t) \to \overline{Q}_\beta(0).$$

Now use (7.30), (7.36) and (7.37) to identify the u.o.c. limit of $\overline{B}^n$ as $\overline{B}^h$ defined in (7.31), identical to the limit of $\overline{B}^{h,n}$. Consequently,

$$(7.38) \qquad \overline{B}^{l,n} \to 0 \quad \text{u.o.c. as } n \to \infty.$$

7.6. We are now ready to prove Theorem 7.1. The convergence of $\hat{Q}^{h,n}$ and $\hat{B}^{h,n}$ and the relations (7.2)–(7.9) are all consequences of Theorem 4.1, so the focus is on l-customers. The first step, as in Proposition 4.4 [cf. (7.38)], is to verify that

$$7.6.A \quad \bar{\bar{B}}^{l,n}(t) = \frac{1}{n}\bar{B}^{l,n}(nt) = \frac{1}{n^2}B^{l,n}(n^2t) \to 0 \quad \text{u.o.c. as } n \to \infty.$$

Since $\bar{\bar{B}}^{l,n}(t)$ is uniformly Lipshitz, 7.6.A will follow from Arzela–Ascoli's theorem once it is shown that any u.o.c. convergent subsequence of $\bar{\bar{B}}^{l,n}(t)$ in fact converges u.o.c. to zero. For ease of notation, let us assume that the sequence $\bar{\bar{B}}^{l,n}$ itself converges u.o.c. and denote its limit by $\bar{\bar{B}}^{l}$. The proof of 7.6.A now amounts to verifying that $\bar{\bar{B}}^{l} \equiv 0$. The convergence of $\bar{\bar{B}}^{l,n}$ implies that $\hat{\xi}^{l,n} \to 0$, in view of (7.29). Consequently,

$$(7.39) \qquad \hat{\xi}^n \to \hat{\xi} \quad \text{u.o.c. as } n \to \infty,$$

where $\hat{\xi}^n$ is defined in (7.28) and $\hat{\xi}$ is a continuous process.

In the derivation of the fluid approximation it was already noted that $\mu_\beta = 0$, hence Theorem 4.1 cannot be applied directly to 7.4.C. Again, we resort to the block representation (7.34)–(7.35). Momentarily assume that

$$7.6.B \qquad \hat{Y}_\beta^n \to Y \quad \text{u.o.c. as } n \to \infty.$$

The representation (7.34), treated as arising from the open network $(0, P_\alpha(l), \mu_\alpha)$, confirms with Theorem 6.1 that

$$(7.40) \qquad \hat{Q}_\alpha^n \to \hat{Q}_\alpha = 0,$$

in accordance with (7.10). Solving for $\hat{Y}_\alpha^n$ in (7.34) and substituting the outcome into (7.35) yields

$$(7.41) \qquad \hat{Q}_\beta^n = \hat{X}_\beta^n + \left[I - \tilde{P}'(l)\right]\hat{Y}_\beta^n,$$

where

$$\hat{X}_\beta^n(t) = \hat{Q}_\beta^n(0) + P'_{\alpha\beta}(l)\left[I - P'_\alpha(l)\right]^{-1}\hat{Q}_\alpha^n(0)$$
$$+ \hat{\xi}_\beta^n(t) + P'_{\alpha\beta}(l)\left[I - P'_\alpha(l)\right]^{-1}\hat{\xi}_\alpha^n(t) - P'_{\alpha\beta}(l)\left[I - P'_\alpha(l)\right]^{-1}\hat{Q}_\alpha^n(t),$$

and $\tilde{P}(l)$ is defined in (7.15). Applying to (7.41) the continuity 3.3.E of the oblique reflection mapping now verifies the convergences of $\hat{Q}_\beta^n$, $\hat{Y}_\beta^n$ and $\hat{B}_\beta^n$ [the latter in view of (7.24)], which proves (7.11)–(7.16) and (7.18). Letting now $n \to \infty$ in the "hat" version of (7.34) establishes the convergence of B_α^n and hence (7.17). Finally, the limits of $\hat{Q}^{l,n}$ and $\hat{B}^{l,n}$ in (7.19) and (7.20) follow from the relations

$$\hat{Q}^{l,n} = \hat{Q}^n - \hat{Q}^{h,n} \quad \text{and} \quad \hat{B}^{l,n} = \hat{B}^n - \hat{B}^{h,n}.$$

To complete the proof of Theorem 7.1, one must still check 7.6.B and show that $\bar{\bar{B}}^{l} \equiv 0$. First, it has been actually shown that the limit of every u.o.c. convergent subsequence of $\{\hat{Y}_\beta^n,\ n \geq 1\}$ must coincide with Y in (7.16). It suffices, therefore, to exhibit one such subsequence. To this end, note that for any $t \geq s \geq 0$,

$$0 \leq \hat{Y}_\beta^n(t) - \hat{Y}_\beta^n(s) \leq -\Delta_\beta^l \left[\hat{B}_\beta^{h,n}(t) - \hat{B}_\beta^{h,n}(s) \right].$$

Since $\hat{B}_\beta^{h,n}$ converges u.o.c. to a continuous limit, the sequence $\{\hat{Y}_\beta^n, n \geq 1\}$ is equicontinuous. The theorem of Arzela–Ascoli now guarantees 7.6.B.

The convergence to $\bar{\bar{B}}^l$ in 7.6.A implies that $\bar{\bar{Y}}_\beta^n$ converges u.o.c. to 0 and hence $\bar{\bar{B}}_\beta^n$ converges u.o.c. to t. The latter convergence, combined with 7.5.A and (7.31), shows that $\bar{\bar{B}}_\beta^l \equiv 0$. Now the block representation (7.34) and the convergence 7.5.A yields $\bar{\bar{B}}_\alpha^l \equiv 0$, thus completing the proof of 7.6.A. Finally note that, in view of 7.6.A, $\hat{\xi}$ in (7.39) is as defined in (7.14).

8. Simple extensions, related results and future research.

EXTENSIONS. The scope of Theorems 4.1, 6.1 and 7.1 can be extended significantly with minor effort. We now outline some possibilities.

8.1. Consider a sequence of networks $(\lambda^{0,n}, P(n), \mu^n)$ in which the switching matrix also varies with n. This additional flexibility manifests itself only through the Brownian drift of the diffusion limits. Specifically, for closed networks add to 4.2.A–4.2.B the assumption that

$$n[P(n) - P] \to M \quad \text{as } n \to \infty, \quad \text{where } P \text{ is stochastic irreducible},$$

and for the open networks add to 6.1.A–6.1.C the convergence

$$\sqrt{n}\,[P(n) - P] \to M \quad \text{as } n \to \infty, \quad \text{where } \sigma(P) < 1.$$

In both cases M is an arbitrary $J \times J$ matrix. Then Theorems 4.1 and 6.1 still prevail with $\hat{\xi} = \mathrm{BM}(M'\mu, \Lambda)$ (instead of $\mathrm{BM}(0, \Lambda)$).

8.2. For the convergence of queue lengths, workloads and busy times, only 4.2.A–4.2.B and 4.5.B–4.5.D are essential in Theorem 4.1 and only 6.1.A–6.1.C and 6.2.A–6.2.D in Theorem 6.1. Our results, therefore, accommodate any arrival, service or routing scheme, as long as the parameters involved converge at the proper rates and the underlying primitive processes jointly satisfy an appropriate FCLT. Such extensions again affect only the parameters of the Brownian motion $\hat{\xi}$. They apply, for example, to batch arrivals, to arrival and service processes that are either superpositions or splittings of renewal processes and to some networks with dependencies among routing, arrivals and services [as in the single-station model analyzed by Fendick, Saksena and Whitt (1988)]. Readers are referred to Section 6 in Reiman (1984) for more rigorous details.

Reiman (1984) also points out that the convergence of sojourn times requires more than the assumptions referred to in the previous paragraph. For concreteness consider closed networks. First, the representation (5.23) is valid only if a FIFO service discipline is adhered to at all stations. Now with FIFO, the convergence of $D_{j,h}^n$ in Theorem 4.1 indeed holds under the assumptions mentioned above *if* 5.6.A prevails, but 5.6.A may fail without Markovian routing.

8.3. Theorem 4.1, as well as Subsections 8.1–8.2, extend to reducible closed networks. Here the switching matrix P is of the form

$$P = \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1r} \\ 0 & P_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_{rr} \end{pmatrix},$$

where the diagonal blocks are all squares, P_{11} is substochastic with $\sigma(P_{11}) < 1$ and P_{ii}, $i = 2, \dots, r$, are irreducible stochastic [see, e.g., Berman and Plemmons (1979), pages 222–223]. Write $\hat{Q}^n$ in Theorem 4.1 as $(\hat{Q}^{1,n}, \dots, \hat{Q}^{r,n})$ according to the partition of P. Then the diffusion limit $\hat{Q}^1(t)$ vanishes for $t > 0$. Each other block i, $i = 2, \dots, r$, behaves like an autonomous irreducible closed subnetwork to which Theorem 4.1 applies, but with the ith block of $\hat{Q}(0)$ in 4.2.A replaced with $\hat{Q}^i(0) + P'_{1i}[I - P'_{11}]^{-1}\hat{Q}^1(0)$.

8.4. Prioritized *open* networks can also be analyzed. As in Theorem 7.1, assume that there are only h- and l-type customers, but here $\sigma[P(h)] < 1$ and $\sigma[P(l)] < 1$. Johnson (1983) considered the case where none of the stations are h-bottlenecks. As expected from (6.2) and Theorem 7.1 then, only the diffusion limits associated with l-customers are nontrivial. They can be calculated as though the network is the single class network from Section 6, but the parameters that determine the limits must be modified to account for the presence of h-customers. Specifically, one uses the service-rate vector $\operatorname{diag}(e - \rho^h)\mu^l$, which discounts the capacity allocated to h-customers and a covariance matrix that is modified similarly to 7.2.D. Suppose, on the other hand, that a certain station j is a bottleneck for h-customers. Then the diffusion limit $\hat{Q}_j^h$ does not vanish. The behaviour of $\hat{Q}_j^l$ conforms to Theorem 7.1 with an additional wrinkle: If $\rho_j^h > 1$, then l-customers accumulate at j at a rate λ_j^l and their effective service rate there is zero.

RELATED RESULTS. We now survey some recent results of others that pertain to the ones reported here.

8.5. It is commonly accepted that Baskett, Chandy, Muntz and Palcois (1975) and Kelly (1979) have come close to exhausting the models of queueing networks (of the type discussed here) which are amenable to exact analysis. This explains the recent surge in research, which our results contribute to and which resorts to approximations as an alternative mode of analysis. Specifically, Theorems 4.1, 6.1 and 7.1 provide theoretical justifications and guidelines for approximating queueing networks in terms of their bottleneck subnetworks. Such approximations, commonly referred to as heavy-traffic approximations, have traditionally treated *balanced* networks, thus excluding networks with either nonbottleneck or strict bottleneck stations. Coffman and Reiman (1984) and Harrison and Williams (1987) approximate balanced open networks, while Harrison, Williams and Chen (1990) focus on balanced closed networks. The theoretical justifications for excluding nonbottlenecks in heavy traffic approximations are (4.2) and (6.2) [the latter was partially derived by Johnson (1983)]; strict bottlenecks have been omitted presumably because they are perceived as rare in real systems. Reiman (1987) is recommended for a summary of the principles that underly heavy traffic approximations, as well as for references to alternative approximation schemes of queueing networks.

8.6. The analysis of the stationary distribution of a stochastic network ranks high in importance. As a first step, one must address the foundational question concerning the existence and uniqueness of such a distribution. A safe conjecture is that, in great generality for the models in Section 2, a stationary distribution exists for closed networks, as well as for open networks without bottlenecks. Major progress toward confirming this conjecture is reported in Borovkov (1987) [see also Kaspi and Mandelbaum (1989)], but a definitive resolution still seems unavailable. Borovkov (1987) also considers some diffusion limits of closed networks. However, his results are not as explicit and are less general than ours and his approach seems less successful than the one employed here.

8.7. Consider an RBM which closely approximates some queueing network and suppose that this RBM has a stationary distribution. Kingman (1965) and Harrison (1973) [see also pages 244–247 in Ethier and Kurtz (1986)] support the hope that the queueing network itself must have a stationary distribution. This distribution, properly normalized, must also be close to that of the RBM, but a formal justification is available only in the easy case of closed networks [Kaspi and Mandelbaum (1989)]. The premise, however, was sufficient to stimulate the works by Harrison and Williams (1987) and Harrison, Williams and Chen (1990). In the first paper, the authors demonstrate that for balanced open networks ($\alpha = \gamma = \varnothing$ in Theorem 6.1), the open RBM $\hat{Q}$ has a unique stationary distribution if and only if

$$c = c^{\mu} - [I - P']^{-1} c^{\lambda} > 0, \tag{8.1}$$

where c^{μ} and c^{λ} are defined in 6.1.A and 6.1.B, respectively. For the sequence of open networks converging to $\hat{Q}$, (8.1) implies that all the traffic intensities of the nth network, n large enough, are indeed strictly less than unity. In the second paper previously mentioned, it is verified for balanced irreducible closed networks that the closed RBM $\hat{Q}$ in Theorem 4.1 always has a unique stationary distribution. (This is expected due to the compact state space involved.)

Explicit calculations of the stationary distribution of an RBM are currently available only when the distribution has a separable density of an exponential product form (an example will be given momentarily). Indeed, the authors mentioned in the previous paragraph prove, both for open and irreducible closed RBM's, that such a form prevails if and only if the covariance matrices $\hat{\Lambda}$ in 4.2.D and 6.1.E satisfy

$$(8.2)\quad 2\hat{\Lambda}_{jk} = -\left[\hat{\Lambda}_{jj}p_{jk}/(1-p_{jj}) + \hat{\Lambda}_{kk}p_{kj}/(1-p_{kk})\right] \quad \text{for all } j \neq k.$$

For an open RBM $\hat{Q} = (\hat{Q}_1, \dots, \hat{Q}_J)'$, (8.2) implies that, at a stationary state, its components are independent random variables and that each $\hat{Q}_j$, $j = 1, \dots, J$, is exponentially distributed with mean $2\mu_j c_j/\hat{\Lambda}_{jj}$, c defined in (8.1). The analogous description for closed RBM's, given by (2.17) in Harrison, Williams and Chen (1990), is omitted here because it is less straightforward to state.

Easy algebra shows that $\hat{\Lambda}$ in Theorems 4.1 and 6.1 satisfies (8.2), respectively, when $b_j = 1$ and when $a_j = b_j = 1$, $j = 1, \dots, J$. (Recall that a_j and b_j are the coefficients of variation of the exogenous interarrival times and the service times at station j.) These latter conditions are clearly met by Poisson arrivals and exponential service times, as in all the classical references listed in Subsection 1.1.

It was already emphasized that the balanced subnetwork β of a general network behaves like an autonomous system whose parameters are identified by our limit theorems. The stationary distribution of this subnetwork, when it exists, has a separable form if $b_\beta = e$ for closed networks and $a_\beta = b_\beta = e$ for open networks. Let us conclude this digression on stationary distributions with an observation that applies to the example in Subsection 1.5: if the service rates μ^n do not vary with n in Theorem 4.1 and (8.2) prevails, then the stationary distribution of the closed RBM that approximates the bottleneck subnetwork β is, in fact, the *uniform* distribution on the unit simplex of dimension $|\beta|$.

8.8. One should mention another research trend that concerns diffusion approximations of stochastic networks. It is aimed at models in which some form of exponentiality is presumed. Such assumptions give rise to Markov and point processes that are analyzable by martingale-based techniques. Representative references are Yamada (1988) and Kogan and Krichagina (1988).

8.9. *Directions for future research.*

NETWORKS WITH FINITE BUFFERS. Kogan and Krichagina (1988), as well as the other papers in Perros and Altiok (1988), are actually concerned with models of networks in which an upper bound is imposed on the number of customers that can simultaneously occupy some of the stations. These practically important models are referred to in the queueing literature as either queueing networks with finite buffers, or with finite capacity, or with blocking; viewed as particle systems [Liggett (1987)], they are called systems with exclusions. (Blocking and exclusion refer to the fact that transitions of customers into fully occupied stations are forbidden.) Fluid and diffusion approximations for networks with finite buffers should arise from increasing the buffer sizes indefinitely and at proper rates. Indeed, order n is probably required for the closed model in Section 4 and order $\sqrt{n}$ for the open model in Section 6. Consequently, such approximations would provide insight on queueing networks with buffers of moderate sizes but, to the best of our knowledge, no theory has yet been developed. We believe that some modification of our approach should be applicable to approximate the finite-buffer versions of the models in Sections 4 and 6. The corresponding fluid and diffusion approximations would arise within the framework for RBM's in polyhedral domains [as in Williams (1987) or Mandelbaum (1990)].

MULTITYPE NETWORKS. The model analyzed in Section 7 is a simple special case of a multitype or multiclass queueing network [Baskett, Chandy, Muntz and Palcois (1975), Kelly (1979)]. These are queueing networks in which the customers are of several types, or classes, and they are allowed to change their types upon completion of each service. [Work on multitype particle systems has only recently started to appear, but with different emphasis; see Durrett and Swindle (1988).] Fluid and diffusion approximations for multitype networks is an active challenging area of research, as demonstrated in Johnson (1983), Peterson (1985), Harrison (1988), Reiman (1987, 1988), Whitt (1988) and Chen and Mandelbaum (1988).

Once customers are distinguishable, there is freedom in specifying the order in which they are served and the routing which they are to follow. Here, different options typically give rise to different fluid and diffusion approximations. This is important because it raises the possibility of comparisons among different operating schemes, perhaps even leading ultimately to those which, under some circumstances, are optimal in some asymptotic sense [cf. Wein (1987), Chen and Yao (1989) and Chen (1990)]. We refer the reader to Harrison (1988), where a promising general framework, called a Brownian network, is introduced. Approximating multiclass networks, however, is typically difficult. Let us conclude with three examples for which some progress has been made, but a definite form has not been reached yet. First consider the prioritized networks in Section 7. Our methods and results are no longer applicable if customers are allowed to change their type upon completion of each service. A second example is when the service discipline at each station is FIFO, but it is required to keep separate track of the performance measures associated with each class. A last example is when each class can be served only by a restricted set of servers (networks with multiserver stations can be accommodated within this framework). An example is when a customer, upon completion of service, joins the server who is confronted by the least work among all the servers by which he can be served.

Acknowledgments. Thanks are due to Mike Harrison, Haya Kaspi, Ward Whitt, Ruth Williams and an anonymous referee for commenting on earlier versions of the manuscript and offering suggestions that have led to its much improved present form.

REFERENCES

ANTHONY, R. N. (1965). *Planning and Control Systems: A Framework for Analysis*. Harvard Univ. Press.

BASKETT, F., CHANDY, K. M., MUNTZ, R. R. and PALACOIS, F. G. (1975). Open, closed and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* **22** 248-260.

BERMAN, A. and PLEMMONS, R. J. (1979). *Nonnegative Matrices in the Mathematical Sciences*. Academic, New York.

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

BOROVKOV, A. A. (1987). Limit theorems for queueing networks, I and II. *Theory Probab. Appl.* **31** 413-427; **32** 257-272 (English translation).

CHEN, H. (1990). Optimal routing control of fluid models. Unpublished manuscript.

CHEN, H. and MANDELBAUM, A. (1988). Open heterogeneous fluid networks. Unpublished manuscript.

CHEN, H. and MANDELBAUM, A. (1991a). Discrete flow networks: Bottleneck analysis and fluid approximations. *Math. Oper. Res.* **16** 408-446.

CHEN, H. and MANDELBAUM, A. (1991b). Leontief systems, RBV's and RBM's. In *Proc. Imperial College Workshop on Applied Stochastic Processes* (M. H. A. Davis and R. J. Elliott, eds.). Gordon and Breach, New York.

CHEN, H. and YAO, D. D. (1989). Optimal scheduling control in a multi-class fluid network. Unpublished manuscript.

COFFMAN, E. G. and REIMAN, M. I. (1984). Diffusion approximations for computer/communication systems. In *Mathematical Computer Performance and Reliability* (G. Iazeolla, P. J. Courtois and A. Hordijk, eds.) 33-53. North-Holland, Amsterdam.

DURRETT, R. and SWINDLE, G. (1988). Are there bushes in a forest? Unpublished manuscript.

ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence*, Wiley, New York.

FENDICK, K. W., SAKSENA, V. R. and WHITT, W. (1988). Dependence in packet queues. *IEEE Trans. Comm.* To appear.

FLORES, C. (1985). Diffusion approximations for computer communications networks. In *Computer Communications. Proc. Symp. Appl. Math.* (B. Gopinath, ed.) 83-124. Amer. Math. Society, Providence, R.I.

GLYNN, P. W. and WHITT, W. (1986). A central-limit-theorem version of $L = \lambda W$. *Queueing Systems* **2** 191-215.

GOODMAN, J. B. and MASSEY, W. A. (1984). The non-ergodic Jackson network. *J. Appl. Probab.* **21** 860-869.

GORDON, W. J. and NEWELL, G. F. (1967). Closed queueing systems with exponential servers. *Oper. Res.* **15** 254-265.

HARRISON, J. M. (1973). The heavy traffic approximation for single server queues in series. *J. Appl. Probab.* **10** 613-629.

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

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

HARRISON, J . M. and REIMAN, M. I. (1981a]. Reflected brownian motion on an orthant. *Ann. Probab.* **9** 302-308.

HARRISON, J. M. and REIMAN, M. I. (1981b). On the distribution of multi-dimensional reflected brownian motion. *SIAM J. Appl. Math.* **41** 345-361.

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

HARRISON, J. M., WILLIAMS, R. and CHEN, H. (1990). Brownian models of closed queueing networks. *Stochastics and Stochastic Reports* **29** 37-74.

IGLEHART D. L. and WHITT, W. (1970). Multiple channel queues in heavy traffic, I and II. *Adv. in Appl. Probab.* **2** 150-177, 355-364.

JACKSON, J. R. (1963). Jobshop-like queueing systems. *Management Sci.* **10** 131–142.

JOHNSON, D. P. (1983). Diffusion approximations for optimal filtering of jump processes and for queueing networks. Ph.D. dissertation, Univ. Wisconsin.

KASPI, H. and MANDELBAUM, A. (1989). On the ergodicity of a closed queueing network. Unpublished manuscript.

KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, New York.

KELLY, F. P. (1984). The dependence of sojourn times in closed queueing networks. In *Mathematical Computer Performance and Reliability* (G. Iazeolla, P. J. Courtois and A. Hordijk, eds.) 111–121. North-Holland, Amsterdam.

KINGMAN, J. F. C. (1965). The heavy traffic approximation in the theory of queues. In *Proc. Symp. Congestion Theory* (W. L. Smith and W. E. Wilkinson, eds.) 137–159. Univ. North Carolina Press, Chapel Hill.

KLEINROCK, L. (1976). *Queueing Systems* II*: Computer Applications*. Wiley, New York.

KOGAN, Y. A. and KRICHAGINA, E. V. (1989). Closed exponential queueing networks with blocking in heavy traffic. In *Proc. Workshop Queueing Networks with Blocking* (H. G. Perros and T. Altiok, eds.) 217–226. North-Holland, Amsterdam.

LEMOINE, A. J. (1978). "Network of Queues—A survey of weak convergence results. *Management Sci.* **24** 1175–1193.

LIGGETT, T. M. (1987). *Interacting Particle Systems*. Springer, New York.

MANDELBAUM, A. (1990). The dynamic complementarity problem. Unpublished manuscript.

MASSEY, B. (1981). Nonstationary queueing networks. Ph.D. dissertation, Stanford Univ.

MOORE, C. G. III (1971). Network models for large-scale time-sharing systems. Technical Report 71–1, Dept. Industrial Engineering, Univ. Michigan.

PERROS, H. G. and ALTIOK, T., EDS. (1988). *Pre-Conf. Proc. of the First International Workshop on Queueing Networks with Blocking*. North-Holland, Amsterdam. To appear.

PETERSON, W. P. (1985). Diffusion approximations for networks of queues with multiple customer types. Ph.D. dissertation, Stanford Univ.

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

REIMAN, M. I. (1982). *The Heavy Traffic Diffusion Approximation for Sojourn Times in Jackson Networks* (R. L. Disney and T. J. Ott, eds.) **2** 409–422. Birkhäuser, Boston.

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

REIMAN, M. I. (1987). A network of priority queues in heavy traffic: One bottleneck station. Unpublished manuscript.

REIMAN, M. I. (1988). A multi-class feedback queue in heavy traffic. *Adv. Appl. Prob.* To appear.

REIMAN, M. I. and WILLIAMS, R. J. (1988). A boundary property of semimartingale reflecting brownian motions. *Probab. Theory Related Fields* **77** 87–97.

RESNICK, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer, New York.

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

SKOROHOD, A. V. (1961). Stochastic differential equations for a bounded region. *Theory Probab. Appl.* **6** 264–274.

SPITZER, F. (1970). Interaction of Markov processes. *Adv. in Math.* **5** 246–290.

STROOCK, D. W. and VARADHAN, S. R. S. (1979). *Multidimensional Diffusion Processes*. Springer, New York.

VARADHAN, S. R. S. and WILLIAMS, R. J. (1985). Brownian motion in a wedge with oblique reflection. *Pure Appl. Math. Sci.* **38** 405–443.

WEIN, L. M. (1987). Asymptotically optimal scheduling of a two-station multi-class queueing network. Ph.D. dissertation, Stanford Univ.

WHITT, W. (1974). Heavy traffic theorems for queues: A survey. *Mathematical Methods in Queueing Theory* (A. B. Clarke, ed.) 307–350. Springer, New York.

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

WHITT, W. (1988). A queueing network analyzer for manufacturing. Unpublished manuscript.

WHITTLE, P. (1967). Nonlinear migration processes. *Bull. Inst. Internat. Statist.* **42** 642–647.

WHITTLE, P. (1968). Equilibrium distributions for an open migration process. *J. Appl. Prob.* **5** 567–571.

WHITTLE, P. (1986). *Systems in Stochastic Equilibrium*. Wiley, New York.

WILLIAMS, R. J. (1987). Reflected brownian motion with skew symmetric data in a polyhedral domain. *Probab. Theory Related Fields* **75** 459–485.

WOLFF, R. W. (1970). Work conserving priorities. *J. Appl. Probab.* **7** 327–337.

WOODS, L. C. (1975). *The Thermodynamics of Fluid Systems*. Oxford Univ. Press.

YAMADA, K. (1988). A heavy traffic limit theorem for $G/M/\infty$ queueing networks. *Probability Theory and Mathematical Statistics. Lecture Notes in Math.* **1299** 549–564. Springer, New York.

FACULTY OF COMMERCE AND BUSINESS ADMINISTRATION
2053 MAIN MALL
VANCOUVER, BRITISH COLUMBIA
CANADA V6T 148

GRADUATE SCHOOL OF BUSINESS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305