The Annals of Applied Probability
1998, Vol. 8, No. 2, 569–646

STATE-DEPENDENT STOCHASTIC NETWORKS. PART I: APPROXIMATIONS AND APPLICATIONS WITH CONTINUOUS DIFFUSION LIMITS

BY AVI MANDELBAUM AND GENNADY PATS

Technion

In a state-dependent queueing network, arrival and service rates, as well as routing probabilities, depend on the vector of queue lengths. For properly normalized such networks, we derive functional laws of large numbers (FLLNs) and functional central limit theorems (FCLTs). The former support fluid approximations and the latter support diffusion refinements.

The fluid limit in FLLN is the unique solution to a multidimensional autonomous ordinary differential equation with state-dependent reflection. The diffusion limit in FCLT is the unique strong solution to a stochastic differential equation with time-dependent reflection.

Examples are provided that demonstrate how such approximations facilitate the design, analysis and optimization of various manufacturing, service, communication and other systems.

Contents

1. Introduction
2. The model of the $(M_\xi/M_\xi/1)^K$ network
3. Martingale representation
4. Fluid approximations (FLLN)
5. An alternative representation of the fluid limit
6. Local traffic intensities
7. Diffusion approximations (FCLT)
8. Approximations for idle-time processes
9. Discontinuous derivatives $\partial\lambda$, $\partial\mu$ and ∂P
10. Applications
    - 10.1. Special regimes of operation
    - 10.2. Special models
    - 10.3. Congestion-dependent dynamics in manufacturing and communication
    - 10.4. Learning systems
    - 10.5. Epidemic models

Received February 1996; revised September 1997.

AMS 1991 *subject classifications.* Primary 60F17, 60G17, 60J70, 60K25, 60K30; secondary 68M20, 90B10, 90B22, 90B30, 90C33.

Key words and phrases. Birth and death process, state-dependent networks, fluid and diffusion approximations, weak convergence, state- and time-dependent oblique reflection, congestion-dependent routing, learning systems, multiserver systems, large finite buffers, transient analysis.

10.6. Data networks with bursty sources
10.7. Multiprocessor systems with breakdowns
10.8. Multiserver systems with breakdowns and blocking
10.9. Stochastic traffic assignment models
10.10. Human-service systems
11. Numerical examples
11.1. Networks with state-dependent routing
11.2. Large finite buffers
11.3. Numerical comparison of different rescaling procedures
12. Proof of the martingale representation
13. Proof of FLLN
13.1. Proofs under Assumptions B
13.2. Proofs under Assumptions A
14. Proof of FCLT
14.1. The main steps
14.2. Proofs
15. Future research
15.1. M_1-convergence
15.2. Time-dependent networks
Appendix A. Projected differential equations
Appendix B. Time-dependent reflection problems
B.1. Formulation of the problem and main properties
B.2. Derivatives of time-dependent reflection operators
Appendix C. Properties of the routing matrix P
Appendix D. Notation

1. Introduction. The paper deals with *state-dependent* open $(M_\xi/M_\xi/1)^K$ queueing networks. These are exponential networks in which arrival and service rates, as well as routing probabilities, depend on the state —the vector of queue lengths. For properly normalized queue-length processes, we derive functional laws of large numbers (FLLNs) and functional central limit theorems (FCLTs). The former support fluid approximations and the latter support diffusion refinements. This paper extends to a network setting our results [59], where the focus is on a single station. Our model for $(M_\xi/M_\xi/1)^K$ is a state-dependent adaptation of that proposed by Massey and Whitt [60] for time-dependent networks.

Stationary analysis of state-dependent networks started with the seminal work of Jackson [37] and culminated in the work of Serfozo [71] (that also includes extensive references). We, on the other hand, are concerned with the *transient evolution* of $(M_\xi/M_\xi/1)^K$ networks. It is typically hard to analyze, yet, it is often important. This is manifested in networks without stationary behavior, such as critically loaded or overloaded networks, networks operating over a finite horizon or exhibiting a periodic evolution, large networks, which often go through long relaxation phases, and more. For a variety of

reasons, exact analysis of our networks, both transient and stationary, is rarely possible. Hence developing different approximating schemes, for example, fluid and diffusion, is a worthwhile undertaking.

A distinguishing feature of our models is state-dependent routing. Here, we extend the work of Krichagina [46] that covered networks with state-*independent* routing. A consequence of this feature is that the characterization of fluid and diffusion limits involves reflection problems with nonconstant directions of reflection, varying with either time or state. Such maps are not as well behaved as the usual multidimensional Skorokhod maps (in particular, they need not be Lipschitz). Hence, we develop new tools to establish convergence, existence and uniqueness of the limits. Our approach to reflection problems with nonconstant directions of reflection is based on Dupuis and Ishii [22] and [23]. (See Appendix B.)

FLLN (Theorem 4.6) holds *in probability*. The fluid limit in FLLN is the unique solution to a multidimensional autonomous ordinary differential equation (DE) with *state-dependent* reflection. As a consequence, the fluid limit of a network is an absolutely continuous function, each coordinate of which can hit and leave zero through time. (This is in contrast to the one-dimensional case [59]; see Figure 2 that displays a periodic orbit.) The proof of FLLN is based on the Lipschitz property of *time-dependent* reflection operator, established in Appendix B. Our original approach was based on differential inclusions, as in [46]. The current treatment, however, is simpler and was inspired by an anonymous referee.

The weak limit given in FCLT (Theorem 7.2) is the unique strong solution to a stochastic differential equation (SDE) with time-dependent reflection. In general, our diffusion limits are Markov processes with discontinuous sample paths and weak convergence is with respect to Skorokhod's M_1-*topology* (see Section 15 for more details). However, in this paper we prove a restricted version of FCLT, which is still very useful in applications. This theorem gives rise to continuous diffusion limits, in which case the convergence is with respect to Skorokhod's J_1-*topology*, which further reduces to *U-convergence* [9]. Here, again, we develop ideas of Krichagina [46], some of which were also anticipated by Anulova [5] (but without proofs). The extension to discontinuous diffusion limits can be found in Pats [65] and will appear in a future paper [64].

The state-dependent model, proposed in Section 2, provides a flexible framework for accommodating a wide variety of phenomena in queueing networks. The results obtained support the design, analysis and optimization of various manufacturing, service, communication and other systems. Some examples, elaborated on in Sections 10 and 11, are manufacturing and computer networks with congestion-dependent routing, services and possibly also various forms of breakdowns [84, 70, 75, 3], learning systems [78], epidemics models [36], traffic assignment models and resource allocation problems [20]. (See Sections 10 and 11 for details.) Moreover, in Section 11 we demonstrate that our approximations are also useful for the analysis of

various networks that do not fit exactly into our framework. In particular, with an appropriate choice of parameters, our limit theorems lead to reasonable approximations for closed networks, networks with large finite buffers, networks governed by shortest-queue routing and more.

Our FLLN and FCLT also unify and generalize existing approximations for many particular models. Examples are state-independent networks [14, 15] and state-dependent overloaded networks [4], both specialized to exponential networks, and some time-dependent networks. (See Sections 10, 11 and 15 for more references and examples.)

The remainder of the paper is organized as follows. In Section 2, we introduce our model of the $(M_\xi/M_\xi/1)^K$ network. Section 3 presents a semimartingale representation of the queueing processes, which is the starting point for our proofs of FLLN and FCLT. In Section 4 we formulate FLLN and provide some guidelines to its proof. In Section 5 we add alternative characterizations of fluid limits. These characterizations support the definition of overloaded, critically loaded and underloaded regimes, provided in Section 6, and they furnish insights into the nature of the fluid limits. The formulation of our FCLT is presented in Section 7, again followed by a proof outline. FLLN and FCLT for idle-time processes are given in Section 8. Refinements are discussed in Section 9. Sections 10 and 11 are devoted to applications of our results. In Section 12, we prove the semimartingale representation of the queueing process. We prove FLLN and FCLT in Sections 13 and 14, respectively. Section 5 contains a proof of the DE characterization of fluid limits. Work in progress, covering discontinuous diffusion limits and time-dependent networks, is motivated in Section 15. The Appendixes provide the technical background for our limit theorems. In particular, Appendix B contains a summary of some new results on the time-dependent reflection problem that appeared elsewhere [57]. Our main notations are summarized in Appendix D.

2. The model of the $(M_\xi/M_\xi/1)^K$ network. We consider an open queueing network that consists of K stations. Each station operates as a single $M_\xi/M_\xi/1$ queue [59]. Transitions of customers between stations are governed by a family of transition probability matrices. A distinguishing feature of our model is that the transition probabilities, as well as the arrival and service rates, depend on the state of the network, namely, the queue lengths at the stations.

Formally, we analyze the $\mathbb{R}_+^K$-valued stochastic *queueing process* $Q = \{Q(t),\ t \geq 0\}$ that satisfies the relations

$$Q(t) = Q(0) + A(t) + F(t) - D(t), \tag{2.1}$$

$$A_k(t) = N_k^+\left(\int_0^t \lambda_k(Q(u))\,du\right), \tag{2.2}$$

$$F_k(t) = \sum_{j=1}^K \int_0^t 1\{U_j[S_j(u)] \in \pi_{jk}(Q(u-))\}\,dD_j(u), \tag{2.3}$$

$$D_k(t) = \int_0^t 1\{Q_k(u-) > 0\}\, dS_k(u), \tag{2.4}$$

$$S_k(t) = N_k^-\left(\int_0^t \mu_k(Q(u))\, du\right), \tag{2.5}$$

where $t \geq 0$, $k = 1, \ldots, K$, $A = (A_1, \ldots, A_K)^T$ and similarly for F and D. [We put $Q(0-) = Q(0)$.] Here, Q is constructed in terms of the following primitives: $Q(0) \in \mathbb{R}_+^K$ is a random vector; $\lambda = \{\lambda_1, \ldots, \lambda_K\}^T$, $\mu = \{\mu_1, \ldots, \mu_K\}^T$: $\mathbb{R}_+^K \to \mathbb{R}_+^K$ are given measurable functions; $\{U_j[l]\}_{l=0}^{\infty}$, $j = 1, \ldots, K$, are sequences of i.i.d. random variables, uniformly distributed on $[0, 1]$; N_k^+, N_k^-, $k = 1, \ldots, K$, are standard (rate 1), right-continuous with left-limits (RCLL) Poisson processes; π_{jk}: $\mathbb{R}_+^K \to \mathscr{E}[0, 1]$, $j, k = 1, \ldots, K$, are measurable (Clarke [16], page 111) set-valued functions, such that

$$\pi_{jk_1}(\xi) \cap \pi_{jk_2}(\xi) = \varnothing \quad \text{for all } k_1 \neq k_2,\ \xi \in \mathbb{R}_+^K. \tag{2.6}$$

All the random quantities in (2.1)–(2.5) are defined on a common complete probability space. Since N_k^+, N_k^-, $k = 1, \ldots, K$, have RCLL sample paths by assumption, we see that Q, A, F, D and S are RCLL as well. (Note that integrals $\int_0^t$ stand for $\int_{[0,t]}$.) A straightforward pathwise construction of (2.1)–(2.5) establishes the existence and uniqueness of Q up to (a possible) *explosion time*. The simplicity of this construction is due to the *pure-jump* character of the primitives. (One could also use the more general argument of Ethier and Kurtz ([25], Theorem 4.1, Chapter 6), through which Q can be defined by a recurrent procedure.)

The quantities involved in the construction have the following interpretation: $Q(0)$ is an initial queue vector; $A = \{A(t),\ t \geq 0\}$ and $F = \{F(t),\ t \geq 0\}$ are counting processes—the kth coordinates $A_k(t)$ and $F_k(t)$ represent the cumulative number of exogenous and endogenous arrivals to station k during $[0, t]$, respectively. Furthermore, $D = \{D(t),\ t \geq 0\}$ and $S = \{S(t),\ t \geq 0\}$ are counting processes—$D_k(t)$ represents the cumulative number of departures from station k during $[0, t]$, whereas $S_k(t)$ counts *potential* departures from station k; this potential is fully realized during intervals $[r, t]$ over which $Q_k(s) > 0$, $s \in [r, t]$. Now, $\lambda(Q) = (\lambda_1(Q), \ldots, \lambda_K(Q))^T$ and $\mu(Q) = (\mu_1(Q), \ldots, \mu_K(Q))^T$ are, respectively, vectors of instantaneous exogenous arrival and service rates at the state Q. Next, let $p_{jk}(\xi)$ denote the Lebesgue measure of $\pi_{jk}(\xi)$, $\xi \in \mathbb{R}_+^K$. From (2.6) it follows that the nonnegative matrix-valued function P: $\mathbb{R}_+^K \to \mathbb{R}_+^{K \times K}$, given by $P(\cdot) = [p_{jk}(\cdot)]_{j,k=1}^K$, has the property that $P(\xi)$ is substochastic for every $\xi \in \mathbb{R}_+^K$; that is,

$$0 \leq \sum_{k=1}^K p_{jk}(\cdot) \leq 1, \qquad j = 1, \ldots, K.$$

Thus, $P(Q)$ is a matrix of instantaneous state-dependent transition probabilities at state Q. Indeed, in view of (2.3), a customer leaving station j at time u (at this moment S_j and D_j both jump) is routed to station k for which

$$U_j\big[S_j(u)\big] \in \pi_{jk}(Q(u-)). \tag{2.7}$$

Given $S_j(u)$ and $Q(u-)$, the event (2.7) has probability $p_{jk}(Q(u-))$, and there exists at most one k for which it prevails. If (2.7) is violated for all k, the customer leaves the network.

MAIN ASSUMPTIONS ON PRIMITIVES.

(M1) The random quantities $Q(0)$, N_k^+, N_k^- and $\{U_k[l]\}_{l=0}^{\infty}$, $k = 1, \ldots, K$, are assumed to be mutually independent.

(M2) Assume that λ, μ satisfy a linear growth constraint. That is, there exists a constant $L_1 > 0$ such that

$$|\lambda(\xi)|_\infty \vee |\mu(\xi)|_\infty \leq L_1(1 + |\xi|), \qquad \xi \in \mathbb{R}_+^K.$$

(M3) Assume that the spectral radii $r(P(\cdot))$ satisfy $\sup_{\xi \in \mathbb{R}_+^K} r(P(\xi)) < 1$.

(M4) Assume that $\mathbf{E}|Q(0)| < \infty$.

REMARK 2.8. The queueing process Q constructed above is a Markov jump process on the K-dimensional nonnegative integer lattice ([25], Theorem 4.1, Chapter 6). It follows from Proposition 13.4 presented below that Assumption (M2) ensures nonexplosion of Q. The sample paths of Q are $\mathbb{R}_+^K$-valued functions, which are RCLL and piecewise constant.

3. Martingale representation. We restate (2.1)–(2.5) in a form that is amenable to analysis. Specifically, (2.1)–(2.5) are equivalent to

$$\begin{aligned} Q(t) = Q(0) &+ \int_0^t \theta(Q(u))\, du + \alpha(t) \\ &+ \int_0^t \left[I - P^T(Q(u))\right] dY(u), \qquad t \geq 0, \end{aligned} \tag{3.1}$$

$$\theta(\cdot) = \lambda(\cdot) + \left[P^T(\cdot) - I\right]\mu(\cdot), \tag{3.2}$$

$$\alpha = M^a + M^f - M^d, \tag{3.3}$$

$$Y(\cdot) = \int_0^{\cdot} I\{Q(u) = 0\}\mu(Q(u))\, du, \tag{3.4}$$

$$M^a = A - \hat{A}, \qquad M^f = F - \hat{F}, \qquad M^d = D - \hat{D}, \tag{3.5}$$

$$\hat{A}(t) = \int_0^t \lambda(Q(u))\, du, \tag{3.6}$$

$$\hat{F}(t) = \int_0^t P^T(Q(u-))\, I\{Q(u-) > 0\}\mu(Q(u))\, du, \tag{3.7}$$

$$\hat{D}(t) = \int_0^t I\{Q(u-) > 0\}\mu(Q(u))\, du. \tag{3.8}$$

Here A, F and D are given by (2.2)–(2.5).

The following technical lemma provides the mathematical framework for our proofs of FLLN and FCLT later on. We shall show that $\hat{A}$, $\hat{F}$ and $\hat{D}$ are

the compensators for A, F and D, respectively (see, e.g., [11], Theorem T8, Chapter 1). The proof of this lemma is postponed to Section 12, as it has no significance for the understanding of later development.

LEMMA 3.9. *Let* $(\Omega, \mathscr{F}, \mathscr{P})$ *denote the common complete probability space on which the random quantities involved in* (2.1)–(2.5) *are defined. Suppose that the Main Assumptions in Section* 2 *are satisfied. Then there exists a filtration* $\mathbf{F}$ *on* $(\Omega, \mathscr{F}, \mathscr{P})$, *satisfying the "usual conditions"* ([40], *page* 10), *such that* M^a, M^f *and* M^d *given by* (3.5) *are vector-valued locally square integrable* ([69], *page* 35) $\mathbf{F}$*-martingales*.

REMARK 3.10. The representation (3.1)–(3.4) of the queueing process has the following interpretation. The function $\theta(Q)$ in (3.1) describes the *potential net flow rate* through the networks, at state Q. Indeed, the coordinate θ_k $(k = 1, \ldots, K)$ is given by

$$\theta_k(Q) = \lambda_k(Q) + \sum_{j=1}^{K} p_{jk}(Q)\mu_j(Q) - \mu_k(Q).$$

Here, the first term on the right-hand side is the rate of exogenous arrivals to station k, the second term is the rate of potential endogenous arrivals to station k from other stations, and the last term is the potential departure rate at station k. The potential is fully realized if none of the stations is idle. The discrepancy between the *real* and *potential* net flow is captured by Y [see (3.4)]. This discrepancy accumulates during idle periods in the network.

Finally, the martingale α in (3.1) encompasses the jumps of the queueing process. As will be seen later, α is negligible on the fluid scale and gives rise to the continuous martingale part on the diffusion scale.

4. Fluid approximations (FLLN). Consider a sequence $(M_\xi^n/M_\xi^n/1)^K$, $n = 1, 2, \ldots,$ of queueing networks, each of which is specified by (2.1)–(2.5) and satisfies the Main Assumptions in Section 2.

A superscript n indicates that the corresponding quantity is related to the nth network. Introduce the rescaled processes $q^n = \{q^n(t),\ t \geq 0\}$, $n = 1, 2, \ldots,$ by

$$q^n(t) = \frac{1}{n} Q^n(t). \tag{4.1}$$

In view of (3.1)–(3.8), q^n has the representation

$$\begin{aligned} q^n(t) = q^n(0) &+ \frac{1}{n}\int_0^t \theta^n(nq^n(u))\, du + \alpha^n(t) \\ &+ \int_0^t \left[I - \left[P^n(nq^n(u))\right]^T\right] dy^n(u), \qquad t \geq 0, \end{aligned} \tag{4.2}$$

$$\theta^n(\cdot) = \lambda^n(\cdot) + \left[\left[P^n(\cdot)\right]^T - I\right]\mu^n(\cdot), \tag{4.3}$$

$$\alpha^n = \frac{1}{n}(M^{a,n} + M^{f,n} - M^{d,n}), \tag{4.4}$$

$$y^n(\cdot) = \frac{1}{n}\int_0^{\cdot} I\{q^n(u) = 0\}\mu^n(nq^n(u))\,du, \tag{4.5}$$

$$M^{a,n} = A^n - \hat{A}^n, \qquad M^{f,n} = F^n - \hat{F}^n, \qquad M^{d,n} = D^n - \hat{D}^n,$$

$$A_k^n(t) = N_k^+\left(\int_0^t \lambda_k^n(nq^n(u))\,du\right),$$

$$F_k^n(t) = \sum_{j=1}^K \int_0^t 1\left\{U_j^n\left(S_j^n(u)\right) \in \pi_{jk}^n(nq^n(u-))\right\} dD_j^n(u),$$

$$D_k^n(t) = \int_0^t 1\{q_k^n(u-) > 0\}\,dS_k^n(u),$$

$$S_k^n(t) = N_k^-\left(\int_0^t \mu_k^n(nq^n(u))\,du\right),$$

$$\hat{A}^n(t) = \int_0^t \lambda^n(nq^n(u))\,du,$$

$$\hat{F}^n(t) = \int_0^t \left[P^n(nq^n(u-))\right]^T I\{q^n(u-) > 0\}\mu^n(nq^n(u))\,du,$$

$$\hat{D}^n(t) = \int_0^t I\{q^n(u-) > 0\}\mu^n(nq^n(u))\,du.$$

We list below the assumptions on the primitives λ^n, μ^n, P^n, and $q^n(0)$, which are used in the formulations and proofs of our theorems.

ASSUMPTIONS A.

(A1) Assume that

$$\frac{1}{n}\lambda^n(n\xi) \to \lambda(\xi), \qquad \frac{1}{n}\mu^n(n\xi) \to \mu(\xi),$$

$$P^n(n\xi) \to P(\xi), \qquad \text{u.o.c.},$$

as $n \uparrow \infty$, where λ, μ and P are given vector- and matrix-valued locally Lipschitz functions and, moreover, $\sup_{\xi \in \mathbb{R}_+^K} r(P(\xi)) < 1$.

(A2) Assume that $|\lambda^n(n\xi)|_\infty \vee |\mu^n(n\xi)|_\infty \le nL_1(1 + |\xi|)$, $\xi \in \mathbb{R}_+^K$, where $n = 1, 2, \ldots$ and L_1 is a given positive constant.

(A3) Assume that $q^n(0) \to_p q(0)$, as $n \uparrow \infty$, where $q(0) \in \mathbb{R}_+^K$ is a given deterministic vector and the sequence $\{\mathbf{E}|q^n(0)|\}$ is bounded uniformly in n.

The asymptotic behavior of $\{q^n\}$ is described by the following theorem, the proof of which is postponed to Section 13.

THEOREM 4.6 (FLLN). *Suppose that Assumptions A are satisfied. Then* $\{q^n\}$ *converges, u.o.c. over* $[0, \infty)$ *in probability, as* $n \uparrow \infty$, *to a deterministic*

absolutely continuous function q. *This* q *is the unique solution to the following DE with state-dependent reflection*:

$$
(4.7)\qquad \begin{cases} q(t) = q(0) + \int_0^t \theta(q(u))\,du \\ \qquad\qquad + \int_0^t \big[I - P^T(q(u))\big]\,dy(u) \geq 0, \qquad t \geq 0, \\ y \text{ is nondecreasing in each coordinate}, \ y(0) = 0, \\ \int_0^\infty 1^T\big[q(t) > 0\big]\,dy(t) = 0, \end{cases}
$$

where

$$
(4.8)\qquad \theta(\cdot) = \lambda(\cdot) + \big[P^T(\cdot) - I\big]\,\mu(\cdot).
$$

In what follows, q will be referred to as the *fluid limit* associated with the network sequence under consideration. To gain insight into the form of q, compare (4.2) with (4.7), in view of Assumptions A: α^n turns out to be negligible and all other terms are easily matched. Existence and uniqueness of the solution to (4.7) follow from [23], due to the Lipschitz properties of λ, μ and P and the uniform boundedness of $r(P)$ [see Assumption (A1)].

REMARK 4.9. The special case of (4.7) when $P \equiv$ const is widely covered in the literature. (See [31], [54], [22] and references therein.) This particular case is known as a differential equation with *oblique reflection* [31].

Recall the following geometric interpretation of $I - P^T$ [54]. The kth column r^k of $I - P^T$, $k = 1, \ldots, K$, is the direction in which q is reflected when it hits the hyperplane $\xi_k = 0$. Moreover, if q hits a point at which several facets $\xi_k = 0$ intersect, then the direction of reflection belongs to some cone. This cone is generated by the corresponding vectors r^k. Thus, in the case $P \equiv$ const, the directions of reflection are constant. By contrast, in (4.7), the directions of reflection vary from point to point on the boundary. In line with this, the reflection problem (4.7) is called *state-dependent*. This is in contrast to *time-dependent* problems in which the directions of reflection are allowed to vary with time only. Time-dependent reflection problems provide the mathematical framework for our FCLT (see Section 7) and Appendix B is devoted to them.

5. An alternative representation of the fluid limit. In this section we provide a characterization of the fluid limit q, defined by (4.7), as the unique solution to a *state-dependent* projected DE (see Appendix A). (For the notion of a projected DE within the context of *normal reflection*, see [6], page 266.) This characterization provides an explicit algorithm for the construction of q and exposes distinctions between fluid approximations for networks and for single stations. (We employ this algorithm in the example at the end of this section, which exhibits a two-station network with a periodic fluid limit.) Moreover, this DE characterization will be used in Section 6 to help introduce the notion of traffic intensities.

From now on, denote $\mathscr{S} \triangleq \mathbb{R}_+^K$. Further, let $N_{\mathscr{S}}(\chi)$ and $T_{\mathscr{S}}(\chi)$ be, respectively, the tangent and normal cones to $\mathscr{S}$ at $\chi \in \mathscr{S}$. (Recall the definitions of the normal and tangent cones from Appendix D.) Clearly,

$$
\begin{aligned}
N_{\mathscr{S}}(\chi) &= \{\zeta \in \mathbb{R}_-^K\colon\ \zeta_k = 0 \text{ if } k \notin \mathscr{I}_0(\chi)\},\\
T_{\mathscr{S}}(\chi) &= \{\zeta \in \mathbb{R}^K\colon\ \zeta_k \geq 0 \text{ whenever } k \in \mathscr{I}_0(\chi)\}.
\end{aligned}
\tag{5.1}
$$

THEOREM 5.2. *The solution q to* (4.7) *is the unique solution to the projected DE*

$$
\dot{q}(\cdot) = \Pi^{\mathscr{T}(q(\cdot))}\{\theta(q(\cdot))\}, \qquad \mathscr{T}(\cdot) = (T_{\mathscr{S}}(\cdot), P(\cdot)),
\tag{5.3}
$$

with the initial condition $q(0)$.

PROOF. In view of the definition of the state-dependent projection Π (Definition A.1), this theorem actually states that there exists a function $\tilde{m}$ such that the following conditions are satisfied for almost every t:

$$
\dot{q}(t) = \theta(q(t)) + \big[I - P^T(q(t))\big]\tilde{m}(t),
\tag{5.4}
$$

$$
\dot{q}(t) \in T_{\mathscr{S}}(q(t)), \qquad -\tilde{m}(t) \in N_{\mathscr{S}}(q(t)),
\tag{5.5}
$$

$$
\dot{q}(t)^T \cdot \tilde{m}(t) = 0.
\tag{5.6}
$$

First, we prove that any solution to (5.4)–(5.6) satisfies (4.7). To this end, let $y(\cdot) = \int_0^{\cdot} \tilde{m}(s)\,ds$. Then the first equation in (4.7) is satisfied. Next, from the second inclusion in (5.5), it follows that y is nondecreasing and the complementarity (last) condition in (4.7) holds. Moreover, since $\dot{q}(t) \in T_{\mathscr{S}}(q(t))$, we obtain that $\dot{q}(t)^T \cdot n \leq 0$ for all $n \in N_{\mathscr{S}}(q(t))$ (for almost every t) and thus $q(\cdot) \in \mathscr{S}$.

Now we show that the solution q to (4.7) satisfies (5.4)–(5.6). Indeed, let $\tilde{m} = \dot{y}$. Evidently, (5.4) and the second inclusion in (5.5) are satisfied. Further, since $q(\cdot) \in \mathscr{S}$, we have that $\dot{q}(t)^T \cdot n = 0$ for all $n \in N_{\mathscr{S}}(q(t))$ (for almost all t). In particular, $\dot{q}(t) \in T_{\mathscr{S}}(q(t))$ and $\dot{q}(t)^T \cdot \tilde{m}(t) = 0$. The proof is thus complete. □

REMARK 5.7. Theorem 5.2 indicates that the fluid limit for a network is a solution to some multidimensional DE with a *discontinuous right-hand side* (cf. [6], page 266). Indeed, it follows from (5.4)–(5.6) that when $q(\cdot) \in \mathscr{S}^0$, the projection Π in (5.3) is an identity mapping. Therefore, $\dot{q}(\cdot) = \theta(\cdot)$. If, at an instant t, q hits $\partial\mathscr{S}$ at a point χ, $\dot{q}(t)$ becomes the oblique projection of $\theta(\chi)$ onto $T_{\mathscr{S}}(\chi)$.

An appropriate framework for investigating DEs with discontinuous right-hand sides [such as (5.3)] is differential inclusions (see [6] and [26]). Krichagina [46] was the first to apply the martingale approach within the context of differential inclusions to derive FLLN for networks with state-*independent* routing.

It was explained in [59] that the fluid limit for a single station is a monotone absolutely continuous function, which absorbs at zero if it ever

reaches it. (The reason for this is that the fluid limit for a single station is closely related to the unique solution of an autonomous first-order ordinary differential equation. It is known that this solution is a strictly monotone or constant function [28], page 40.) The following corollary points to the fact that, for networks, *the origin of the coordinates* is the absorbing point for fluid limits. It is an immediate consequence of the uniqueness of the solution to (5.3).

COROLLARY 5.8. *If* $q(t_0) = 0$ *for some* $t_0 > 0$, *then* $q(t) \equiv 0$ *for all* $t \geq t_0$.

In contrast to a single station, fluid limits for networks are, in general, nonmonotone functions; each coordinate can hit and leave zero. To illustrate this, we present an example with a periodic fluid limit, in which one of the coordinates hits and leaves zero periodically. The example also provides insight into the nature of representations (4.7) and (5.3). This example can be skipped without loss of reading continuity. (A comprehensive analysis of trajectories of fluid limits as solutions to DEs with discontinuous right-hand sides is beyond the scope of our paper. For this issue, refer to the book by Filippov [26], Chapter 4.)

EXAMPLE. Consider the two-station tandem network depicted in Figure 1, with the primitives

$$\lambda_1(Q) = \begin{cases} 6n, & \text{if } 0 \leq Q_1, Q_2 \leq n, \\ 6n + 20(Q_1 - n)^+, & \text{if } Q_1 > n, \\ \left(6n - 5(Q_1 - n)^- \left(\dfrac{Q_2}{n} - 1\right)^+\right)^+, & \text{otherwise,} \end{cases}$$

$$\mu_1(Q) = \begin{cases} 3n, & \text{if } 0 \leq Q_1, Q_2 \leq n, \\ 6n + 20(Q_1 - n)^+ + 4(Q_2 - n)^+, & \text{if } Q_2 > n, \\ 6n + 20(Q_1 - n)^+, & \text{otherwise,} \end{cases}$$

$$\lambda_2 = 0, \qquad \mu_2 = 5n, \qquad P(\cdot) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

for some $n \in \mathscr{Z}_+$. Here, $Q = (Q_1, Q_2)$, where Q_1 and Q_2 are values of the queues at the first and the second station, respectively.

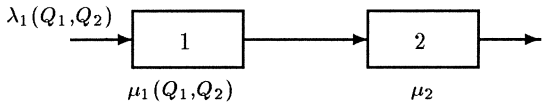


FIG. 1. *A two-station tandem network with a periodic fluid limit.*

By Theorem 4.6, the fluid limit for this network is a solution to (4.7) with

$$\lambda_1(\xi) = \begin{cases} 6, & \text{if } 0 \le \xi_1, \xi_2 \le 1, \\ 6 + 20(\xi_1 - 1)^+, & \text{if } \xi_1 > 1, \\ \left(6 - 5(\xi_1 - 1)^- (\xi_2 - 1)^+\right)^+, & \text{otherwise}, \end{cases} \tag{5.9}$$

$$\mu_1(\xi) = \begin{cases} 3, & \text{if } \xi_1, \xi_2 \le 1, \\ 6 + 20(\xi_1 - 1)^+ + 4(\xi_2 - 1)^+, & \text{if } \xi_2 > 1, \\ 6 + 20(\xi_1 - 1)^+, & \text{otherwise}, \end{cases} \tag{5.10}$$

$$\lambda_2 = 0, \qquad \mu_2 = 5, \qquad P(\cdot) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \tag{5.11}$$

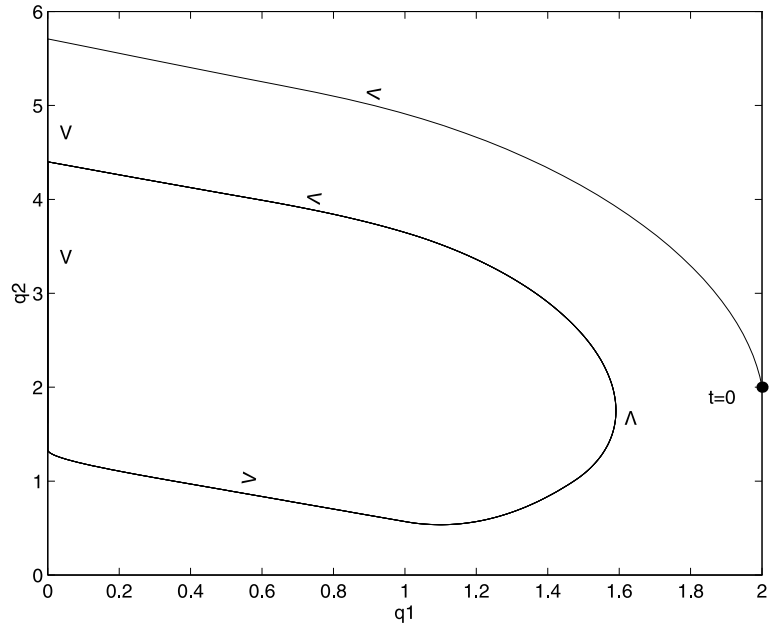
where $\xi_1, \xi_2 \ge 0$. Assume that $Q_1(0) = Q_2(0) = 2n$; that is, $q_1(0) = q_2(0) = 2$. (The form of the fluid limit is sensitive to the initial state. Points other than $Q_1(0) = Q_2(0) = 2n$ need not give rise to periodic orbits.) The fluid limit for this network is depicted in Figure 2a and b. The path goes into a periodic orbit, and q_1, q_2 evolve as periodic functions after the initial transient phase. These graphs are obtained by numerical integration of (5.4)–(5.6).

Constructing the trajectory of the fluid limit. Below, we discuss how the parameters of the network, given by (5.9)–(5.11), give rise to the fluid limit depicted in Figure 2a and b. Substituting (5.9)–(5.11) into (4.8) yields the expression for θ. The vector field generated by θ is illustrated in Figure 2c. This vector field, together with the reflection matrix $[I - P^T]$, defines $\dot{q}$ and, eventually, q. [See (5.4)–(5.6).] Namely, when $q_1, q_2 > 0$, the trajectory of q evolves according to θ (see Remark 5.7). Further, when q hits $\partial\mathbb{R}^2_+$, $[I - P^T]$ comes into play. Specifically, recall from Remark 4.9 that the columns of $[I - P^T]$, $r^1 = [1, -1]'$ and $r^2 = [0, 1]'$ constitute the directions in which q is reflected when it hits the boundary $\chi_1 = 0$ or $\chi_2 = 0$, respectively. For example, consider point in time $t' = 2.24$. At that instant, q hits the boundary $\chi_1 = 0$ at point $[0, 4.4]$. Calculations give that $\theta(0, 4.4) = [-16.4, 11.4]$. Now, (5.4)–(5.6) yield that $g \triangleq [I - P^T(q(t'))]\tilde{m} = [16.4, -16.4]$ and $\dot{q}(t' +) = \theta(q(t')) + g = [0, -5]$. Note that g is colinear to r^1, as it must be. [These calculations are illustrated in Figure 2d.] According to $\dot{q}(t' +)$, q starts moving downward along the axis $\chi_1 = 0$ and keeps this direction until entering the region in which θ points toward the interior of $\mathbb{R}^2_+$. At that instant (around the point $[0, 1.33]$), q leaves the boundary and follows the periodic orbit, as depicted in Figure 2a.

6. Local traffic intensities. This section sets the stage for our FCLT presented in Section 7. We prove additional details on the issue of characterizing traffic intensities in Section 15, within the context of M_1-convergence.

Introduce the function m,

$$\begin{aligned} m(\xi) &= \left[I - P^T(\xi)\right]^{-1}\left(\Pi^{\mathscr{F}(\xi)}\{\theta(\xi)\} - \theta(\xi)\right), \\ \mathscr{F}(\cdot) &= (T_{\mathscr{S}}(\cdot), P(\cdot)), \end{aligned} \tag{6.1}$$



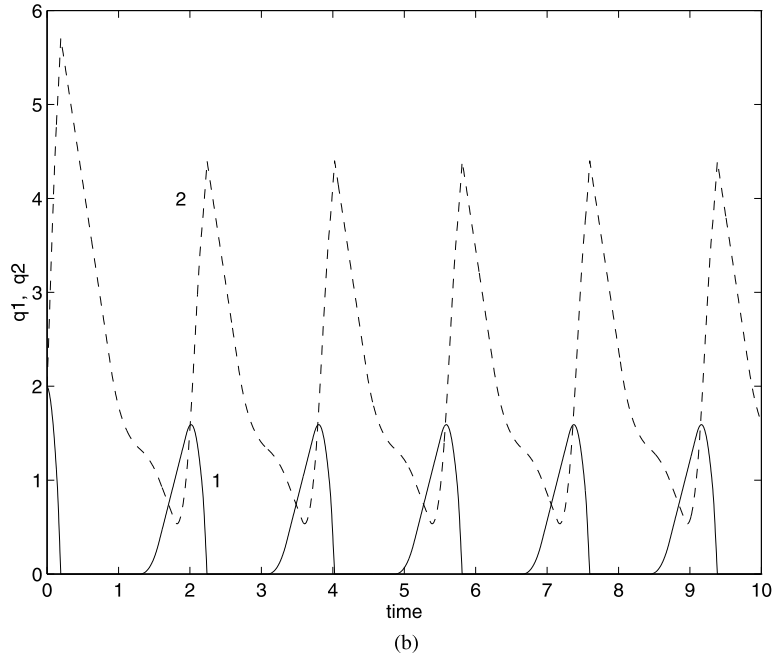
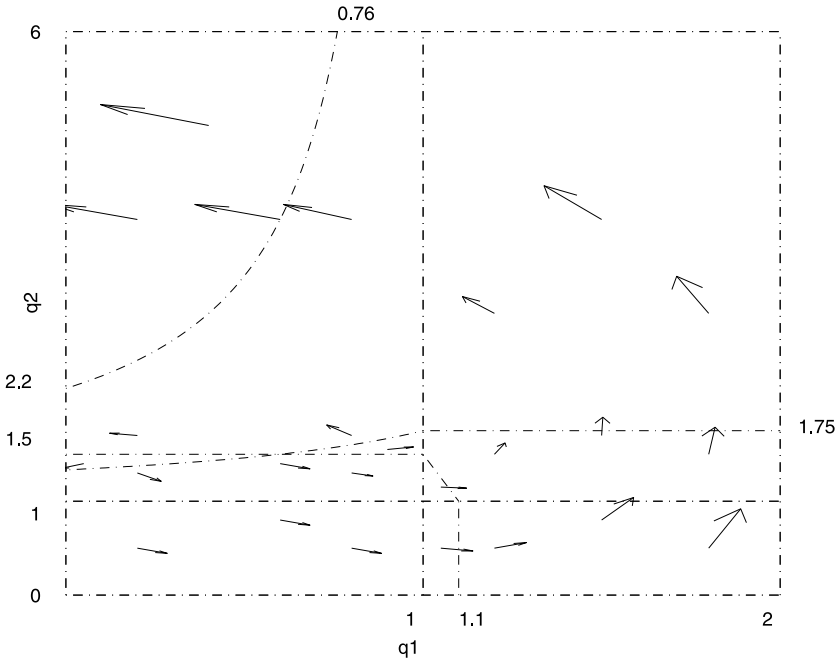
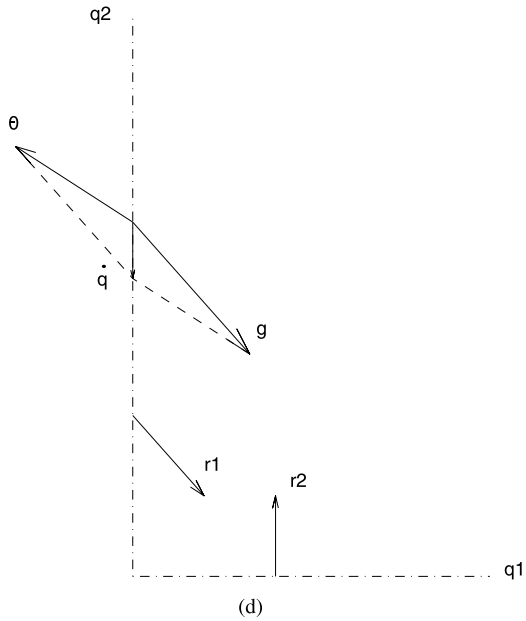


FIG. 2. *Fluid limit for the two-station tandem network*. (a) *Path of the fluid limit*, *which starts from point* $(2, 2)$ *at* $t = 0$ *and follows the arrows*. (b) *Trajectories of the fluid limit*. *The solid line is* q_1, *the dashed line is* q_2. (c) *The vector field generated by* θ. *The broken lines divide the plane into the different regions*, *which are computed by* (4.8) *and* (5.9)–(5.11). (d) *Oblique reflection*: *relative position of* $\theta(q(t'))$, $g = [I - P^T(q(t'))]\dot{y}(t' +)$ *and* $\dot{q}(t' +)$ *at time* $t' = 2.3$.



(c)



(d)

FIG. 2. *Continued.*

where $\xi \in \mathbb{R}_+^K$. In view of Theorems 4.6 and 5.2,

$$m(q(t)) = \dot{y}(t) = \tilde{m}(t), \quad \text{for almost every } t > 0,$$

where y and $\tilde{m}$ are defined by (4.7) and (5.4)–(5.6), respectively. Henceforth, we assume that $\tilde{m}$ is defined for all $t \in [0, T]$ and is given by $\tilde{m}(\cdot) = m(q(\cdot))$.

At each moment t, define the sets $J^+(t)$, $J^-(t)$ and $J^0(t)$ of overloaded, underloaded and critically loaded stations, respectively (cf. [46], [14] and [55]) by

$$\begin{aligned} J^+(t) &= \{j\colon q_j(t) > 0\}, \\ J^-(t) &= \{j\colon q_j(t) = 0,\ m_j(q(t)) > 0\}, \\ J^0(t) &= \{j\colon q_j(t) = 0,\ m_j(q(t)) = 0\}. \end{aligned} \tag{6.2}$$

During its evolution, each station of the network can alternate between overloaded, underloaded and critically loaded phases. To underline that these phases are determined by the fluid limit, we also refer to them as overloaded, critically loaded and underloaded *asymptotic* regions (cf. [55]). Our FCLT presented in the next station reveals that each asymptotic region has its distinctive type of diffusion limit. Specifically, the diffusion limits for overloaded, critically loaded and underloaded stations are diffusion, reflected diffusion and zero processes, respectively.

7. Diffusion approximations (FCLT). Introduce the sequence of stochastic processes $V^n = \{V^n(t),\ t \geq 0\}$, $n = 1, 2, \ldots,$ by

$$V^n(t) = \sqrt{n}\,(q^n(t) - q(t)). \tag{7.1}$$

This sequence represents amplified deviations of the rescaled queueing processes q^n from their fluid limit q. The asymptotic behavior of $\{V^n\}$ is given by FCLT that will be presented momentarily.

Let us start with a brief discussion on the issue of continuity of diffusion limits. Simple analysis of (7.1) reveals that our diffusion limits could, in general, be discontinuous. Indeed, recall the definition (6.2) of $J^+(t)$, $J^-(t)$ and $J^0(t)$. If, for example, $k \in J^+$ for $t < t_0$ and $k \in J^-$ for $t \geq t_0$ (for some $t_0 > 0$), then the limit of $\{V_k^n\}$ has a jump at t_0, with positive probability. Since the jumps of V^n are of size $1/\sqrt{n}$, $\{V_k^n\}$ cannot converge in the usual Skorokhod J_1-topology (the "largest jump" functional is J_1-continuous; see, e.g., [67]). We assert that, in fact, the convergence holds in the Skorokhod M_1-topology. (See Section 15 for a more detailed discussion on this issue.)

In this paper, we prove a simplified version of FCLT, which is yet very useful in applications. Namely, Theorem 7.2 is formulated for an interval $[0, T]$, over which J^+, J^- and J^0 do not depend on time. In this case the limit process is continuous and the convergence holds with respect to the J_1-topology. (Actually, the convergence to a continuous limit holds with respect to the U-topology [9].) Such a version, though simplified, still gives rise to the general form of the diffusion limits, while avoiding the issue of convergence near discontinuity points of the limit. General versions of FCLT,

covering cases of discontinuous limits, appear in Pats [65] and will be presented in a future paper [64]. (The main idea of those versions is to divide the time interval $[0,T]$ into open subintervals, each of which does not contain points of discontinuities, and to apply Theorem 7.2 separately to each of the subintervals. The diffusion processes thus obtained are joined together in a way that possibly gives rise to discontinuities of the final diffusion limit.)

THEOREM 7.2 (FCLT). *Let the conditions of FLLN* (*Theorem* 4.6) *be satisfied. Assume in addition that*

$$\sqrt{n}\left(\frac{\lambda^n(n\xi)}{n}-\lambda(\xi)\right)\to f_\lambda(\xi),\qquad \sqrt{n}\left(\frac{\mu^n(n\xi)}{n}-\mu(\xi)\right)\to f_\mu(\xi),$$
$$\sqrt{n}\,(P^n(n\xi)-P(\xi))\to f_P(\xi),\qquad u.o.c., \tag{7.3}$$

as $n\uparrow\infty$, *where* f_λ, f_μ *and* f_P *are given vector- and matrix-valued functions, which are bounded and continuous. Finally, suppose that*:

(i) λ, μ *and* P *are differentiable with continuous bounded derivatives*;
(ii) $J^+(\cdot)$, $J^-(\cdot)$ *and* $J^0(\cdot)$ *are constant during* $[0,T]$;
(iii) $V^n(0)\to_d V(0)$, *where* $V(0)$ *is a given random vector with* $V_k(0)=0$ *for all* $k\in J^-$.

Then the sequence $\{V^n\}$ *converges weakly over* $[0,T]$ *to a continuous Markov process* V. *The process* V *is the unique* (*strong*) *solution to the SDE with time-dependent reflection* [*see* (B.6)],

$$V=\mathbf{\Phi}_\Gamma^{\mathbf{R}}\left([I-I^-]\int_0^{\cdot}\tilde{R}^{-1}(t)\,dX(t)\right), \tag{7.4}$$

$$\begin{aligned} dX(t) &= f_\theta(q(t))\,dt - f_P^T(q(t))\,dy(t) + \partial\theta(q(t))V(t)\,dt \\ &\quad - \partial P^T(q(t))\odot V(t)\,dy(t) \\ &\quad + \Sigma^{1/2}(q(t))\,dW(t),\qquad t\in[0,T], \end{aligned} \tag{7.5}$$

with the initial condition $X(0)=V(0)$. *Here* q, y *and* θ *are given by* (4.7) *and* (4.8), *respectively*, W *is a standard* $\mathbb{R}^K$*-valued Brownian motion and*

$$\Gamma=\{\xi\in\mathbb{R}^K\colon\ \xi_k\ge 0,\ \forall k\in J^-\cup J^0\}, \tag{7.6}$$

$$R(\cdot)=[I-P^T(q(\cdot))], \tag{7.7}$$

$$\tilde{R}(\cdot)=[I-P^T(q(\cdot))I^-], \tag{7.8}$$

$$f_\theta=f_\lambda+P^Tf_\mu-f_\mu+f_P^T\mu, \tag{7.9}$$

$$\begin{aligned} \Sigma &= \operatorname{diag}\{\lambda\}+\operatorname{diag}\{\mu-m\}+\operatorname{diag}\{P^T(\mu-m)\} \\ &\quad - P^T\operatorname{diag}\{\mu-m\}-\operatorname{diag}\{\mu-m\}P, \end{aligned} \tag{7.10}$$

where m *is given by* (6.1).

In what follows, V will be referred to as the *diffusion limit* associated with the network sequence under consideration.

REMARK. We assume for convenience that $X(0-) = 0$. Hence,

$$V(0) = [I - I^-][I - P^T(q(0))I^-]^{-1}X(0) = X(0),$$

as it must be. (To verify these equalities, recall that all integrals $\int_0^t$ stand for $\int_{[0,t]}$ and see Remark B.3.)

REMARK 7.11. From the definition of $\odot$ (see Appendix D) it follows that the matrix-valued function $E(\cdot) \triangleq \partial P^T(q(\cdot)) \odot V(\cdot)$ satisfies

$$E_{jk}(t) = \sum_{i=1}^{K} \frac{\partial P_{kj}}{\partial \xi_i}(q(t))V_i(t), \qquad j,k = 1,\dots,K,\ t \in [0,T].$$

REMARK. Our FLLN and FCLT can be adapted to cover some cases when λ, μ and P have *piecewise continuous* derivatives. We address this issue in Section 9.

The proof of Theorem 7.2 is postponed to Section 14. Here, we content ourselves with

OUTLINE OF PROOF. In view of (13.2), (13.3) and (7.1), we can write

$$V^n = \sqrt{n}\left[\Phi_{\mathscr{G}}^{\mathbf{R}}\left(x + \frac{1}{\sqrt{n}}X^n\right) - \Phi_{\mathscr{G}}^{\mathbf{R}}(x)\right], \tag{7.12}$$

where

$$X^n = V^n(0) + f_\theta^n - f_P^n + B_\theta^n - B_P^n + M^n, \tag{7.13}$$

$$f_\theta^n(\cdot) = \sqrt{n}\int_0^\cdot \left\{\frac{1}{n}\theta^n(nq^n(u)) - \theta(q^n(u))\right\} du, \tag{7.14}$$

$$f_P^n(\cdot) = \frac{1}{\sqrt{n}}\int_0^\cdot \left[[P^n(nq^n(u-))]^T - P^T(q^n(u))\right] \times I\{q^n(u) = 0\}\mu^n(nq^n(u))\, du, \tag{7.15}$$

$$B_\theta^n(\cdot) = \sqrt{n}\int_0^\cdot \{\theta(q^n(u)) - \theta(q(u))\}\, du, \tag{7.16}$$

$$B_P^n(\cdot) = \frac{1}{\sqrt{n}}\int_0^\cdot \left[P^T(q^n(u)) - P^T(q(u))\right] \times I\{q^n(u) = 0\}\mu^n(nq^n(u))\, du, \tag{7.17}$$

$$M^n = \sqrt{n}\,\alpha^n. \tag{7.18}$$

Straightforward analysis of (7.13)–(7.18) reveals that $\{X^n\}$ is C-tight (see Lemma 14.13). Let X be any weak limit of $\{X^n\}$. Then we may and do assume

that $\{X^n\}$ converges u.o.c., a.s., to X. Next, rewrite V^n in the following way:

$$V^n = \sqrt{n}\left[\Phi_{\mathscr{G}}^{\mathbf{R}}\left(x + \frac{1}{\sqrt{n}}X\right) - \Phi_{\mathscr{G}}^{\mathbf{R}}(x)\right] + \epsilon^n, \tag{7.19}$$

$$\epsilon^n = \sqrt{n}\left[\Phi_{\mathscr{G}}^{\mathbf{R}}\left(x + \frac{1}{\sqrt{n}}X^n\right) - \Phi_{\mathscr{G}}^{\mathbf{R}}\left(x + \frac{1}{\sqrt{n}}X\right)\right]. \tag{7.20}$$

Observe that $\{\epsilon^n\}$ converges to zero u.o.c., a.s. Indeed, in view of the Lipschitz property of time-dependent reflection (see Theorem B.1), we have

$$\|\epsilon^n\|_T \le L\|X^n - X\|_T$$

for some $L > 0$. The limit of $\{V^n\}$ can be interpreted as some form of a directional derivative of $\Phi_{\mathscr{G}}^{\mathbf{R}}$, at the point x in the direction of X. Theorem B.2 provides an expression for this derivative. By that theorem, $\{V^n\}$ converges u.o.c., a.s., to a process V, which is given by (7.4). Actually, it leads to the conclusion that $\{V^n\}$ is C-tight, provided that $\{X^n\}$ is C-tight. To complete the proof of the theorem, we must show, first, that (7.4) and (7.5) possess a unique strong solution (see Lemma 14.8) and, second, that X is given by (7.5) (see Lemma 14.14).

We conclude the outline of the proof with an explanation of the correspondence between (7.5) and (7.13). The first and the second terms on the right-hand side of (7.5) are the limits of $\{f_\theta^n\}$ and $\{f_P^n\}$, respectively. [This is a consequence of (7.3) and FLLN.] Further, applying the mean value theorem and FLLN to (7.16) and (7.17) reveals that $\{B_\theta^n\}$ and $\{B_P^n\}$ give rise in the limit to the third and fourth terms, respectively (see Lemma 14.14). Finally, the last (martingale) term arises from the martingale sequence $\{M^n\}$ (see Lemma 14.9). □

8. Approximations for idle-time processes. A straightforward modification of arguments used in the proofs of FLLN and FCLT (Theorems 4.6 and 7.2, respectively) leads to a corresponding limit theorem for the sequence $\{y^n\}$, given by

$$y^n = \frac{1}{n}Y^n, \qquad Y^n(\cdot) = \int_0^{\cdot} I\{q^n(u) = 0\}\,\mu^n(nq^n(u))\,du, \qquad n = 1, 2, \ldots .$$

This sequence represents a rescaled discrepancy between *real* and *potential* departures, which arises during idle periods in the stations. Note that if at some station k, the service rate depends on the value of queue at that station only [that is, $\mu_k^n(Q^n) = \mu_k^n(Q_k^n)$], then

$$Y_k^n = \mu_k^n(0)\,I_k^n, \tag{8.1}$$

where $I_k^n(\cdot) = \int_0^{\cdot} 1\{q_k^n(u) = 0\}\,du$ is the idle-time process at station k.

The following proposition constitutes the FLLN and FCLT for $\{y^n\}$:

PROPOSITION 8.2. *Assume that the conditions of Theorem* 4.6 *are satisfied. Then* $\{y^n\}$ *converges, u.o.c. over* $[0, \infty)$ *in probability, as* $n \uparrow \infty$, *to a determinis-*

tic absolutely continuous function y given by (4.7). *Assume further that the conditions of Theorem* 7.2 *are satisfied. Then the sequence* $\{H^n\}$, *given by*

$$H^n = \sqrt{n}\,(y^n - y), \qquad n = 1, 2, \ldots,$$

converges weakly to the continuous Markov process

$$H(\cdot) = \int_0^{\cdot} \tilde{R}^{-1}(t)\, d\{V(t) - X(t)\}, \tag{8.3}$$

where V, X *and* $\tilde{R}$ *are characterized by* (7.4), (7.5) *and* (7.8), *respectively*.

9. Discontinuous derivatives $\partial\lambda$, $\partial\mu$ and ∂P. In this section we discuss an extension of FCLT, covering some cases when λ, μ and P have *piecewise continuous derivatives*. In the one-dimensional case, a general statement was given by Theorem 4.3 in [59]. We provided there a condition on the discontinuities of λ and μ, under which FCLT holds without any changes. We also derived a modified FCLT, covering cases when that condition is not satisfied. The case of networks with piecewise continuous derivatives of primitives is treated in [58], where, in particular, queues with reneging, preemptive priorities, finite population and finite number of servers are covered.

In this paper, we present a simple modification of the FCLT which is sufficient for our applications. This version characterizes some cases when FCLT (Theorem 7.2) holds without changes.

PROPOSITION 9.1. *Suppose that all the conditions of Theorem* 7.2 *are satisfied with the following modification*: *there exists* $\varepsilon > 0$ *such that the derivatives* $\partial\lambda$, $\partial\mu$ *and* ∂P *are Lipschitz continuous in the set*

$$\bigcup_{t \in [0,T]} \mathbf{B}[q(t), \varepsilon], \tag{9.2}$$

where $\mathbf{B}[\xi, \varepsilon]$ *is a Euclidean ball with center* ξ *and radius* ε. *Out of this set, we allow* $\partial\lambda$, $\partial\mu$ *and* ∂P *to be piecewise continuous functions with a finite number of discontinuities in each compact subset of* $\mathbb{R}_+^K$. *Then Theorem* 7.2 *applies without any changes*.

The proof is omitted because of its similarity to the proof of Theorem 7.2.

10. Applications. In this section, we demonstrate that our state-dependent networks are natural models of various real systems. Specifically, our examples show that the model (2.1)–(2.5) fits a wide variety of queueing networks, phenomena and forms of control. In most examples, we do not provide explicit expressions for fluid and diffusion limits, due to space limits. In each case, we can write down these expressions by substituting the parameters of the models into the general equations (4.7), (7.4) and (8.3). Note, however, that even for small-size models, such as those considered

below, the corresponding DEs and SDEs often allow only numerical solutions. This motivates the additional numerical examples in Section 11.

10.1. *Special regimes of operation*. We start with particular cases in which the fluid and diffusion limits, given by (4.7), (7.4), (7.5) and (8.3), substantially simplify. These cases will be used in our examples later on in this and the next section.

Networks without underloaded stations. In this case $J^- \equiv \varnothing$ and we have $\dot{q} = \theta(q)$, where θ is given by (4.8). Further, V is the unique (strong) solution to the following SDE with time-dependent reflection:

$$\begin{cases} V(\cdot) = X(\cdot) + \int_0^{\cdot} \big[I - P^T(q(t))\big]\, dY(t), \\ V_k \geq 0, \qquad k \in J^0; \\ Y \text{ is nondecreasing in each coordinate, } Y(0) = 0; \\ Y_k \equiv 0, \qquad k \in J^+; \\ \int_0^{\cdot} 1\{V_k(t) > 0\}\, dY_k(t) \equiv 0, \qquad k \in J^0. \end{cases} \tag{10.1}$$

Here X is defined by

$$dX(t) = f_\theta(q(t))\, dt + \partial\theta(q(t))V(t)\, dt + \Sigma^{1/2}(q(t))\, dW(t), \qquad t \geq 0,$$

and f_θ and Σ are given by (7.9) and (7.10), respectively. Finally,

$$H = Y. \tag{10.2}$$

Overloaded networks. This is a special case of the previous one, in which $J^0 = \varnothing$ (that is, $J^+ \equiv \{1, \ldots, K\}$). Then, again, $\dot{q} = \theta(q)$, (10.1) and (10.2) imply that $H \equiv 0$ and

$$dV(t) = f_\theta(q(t))\, dt + \partial\theta(q(t))V(t)\, dt + \Sigma^{1/2}(q(t))\, dW(t), \qquad t \geq 0.$$

Note that V is a Gaussian process, provided $V(0)$ is a normal random variable, for example, independent of W.

Further, introduce the mean *vector* and covariance *matrix* functions

$$a(\cdot) \triangleq \mathbf{E}V(\cdot), \qquad b(\cdot) \triangleq \operatorname{Cov} V(\cdot) = \mathbf{E}\Big[(V(\cdot) - a(\cdot))(V(\cdot) - a(\cdot))^T\Big].$$

Then (see [40]) a and b satisfy the DEs

$$\begin{aligned} \dot{a}(t) &= f_\theta(q(t)) + \partial\theta(q(t))a(t), \\ \dot{b}(t) &= \partial\theta(q(t))b(t) + b(t)\big[\partial\theta(q(t))\big]^T + \Sigma(q(t)), \qquad t \geq 0. \end{aligned} \tag{10.3}$$

Underloaded networks. In this case, $J^- \equiv \{1, \ldots, K\}$ and we have

$$q \equiv 0, \qquad V \equiv 0;$$

$$H = -\big[I - P^T(0)\big]^{-1}\big\{\big(f_\theta(0) - f_P^T(0)m\big)t + \Sigma^{1/2}(0)W(t)\big\}, \qquad t \geq 0.$$

Here, f_θ, f_P and Σ are the same as in Theorem 7.2 and $m = -[I - P^T(0)]^{-1}\theta(0)$.

10.2. *Special models*. Here is a list of particular models that are covered by our FLLN and FCLT (Theorems 4.6 and 7.2 and Propositions 9.1 and 8.2).

Networks with state-independent routing. For such models, fluid and diffusion limits are solutions to DEs and SDEs, with reflection in *constant* directions. For this case, we extend the results of Krichagina [46] to unbounded arrival and service rates, and we develop a framework for a rigorous analysis of M_1-convergence, to appear in [56] and [64] (see also [59] and Section 15).

Networks with finite population, multiserver stations and state-independent routing. In such models, rates of arrivals and services are given by piecewise linear functions. (See Sections 5.4–5.8 in [59] for single-station examples.) In line with this, fluid limits are solutions to autonomous linear DEs with reflection, while diffusion limits are (reflecting) diffusion processes of the Ornstein–Uhlenbeck type. Our theorems here generalize the corresponding results of Kogan, Liptser and Smorodinskii [45], Prisgrove [68] and Kogan and Liptser [44].

It is of interest that, for some of these networks, our fluid and diffusion approximations provide *exact* expressions for mean values and covariances of the queueing processes. For illustration, consider a sequence of single stations with primitives

$$\lambda^n(Q^n) = \lambda \cdot (n - Q^n), \qquad \mu^n(Q^n) = \mu Q^n, \tag{10.4}$$
$$Q^n(0) = nq(0), \qquad n = 1, 2, \ldots,$$

where $\lambda, \mu \in \mathbb{R}_+$ and $q(0) \in \{0, 1, \ldots, n\}$. By FLLN and FCLT and in view of (10.3) we have

$$q(t) = \frac{\lambda}{\lambda + \mu} - e^{-(\lambda+\mu)t}\left[\frac{\lambda}{\lambda + \mu} - q(0)\right], \qquad a = \mathbf{E}V \equiv 0, \tag{10.5}$$

$$\begin{aligned} b(t) &= \operatorname{Var} V(t) \\ &= \frac{1}{(\lambda + \mu)^2}\left\{\lambda\mu - e^{-(\lambda+\mu)t}\left[\lambda(\mu - \lambda) - q(0)(\mu^2 - \lambda^2)\right]\right. \\ &\qquad \left. - e^{-2(\lambda+\mu)t}\left[\lambda^2 + q(0)(\mu^2 - \lambda^2)\right]\right\}, \qquad t \geq 0. \end{aligned} \tag{10.6}$$

On the other hand, standard calculations with probability generating functions yield

$$\mathbf{E}Q^n = nq, \qquad \operatorname{Var} Q^n = nb, \qquad n = 1, 2, \ldots, \tag{10.7}$$

where q and b are given by (10.5) and (10.6) (see [74] and [27]). Observe that (10.7) is precisely the expression that we obtain by combining (10.5) and (10.6) with the formal relation suggested by FCLT: $Q^n \sim_d nq + \sqrt{n}\, V$. Roughly speaking, the following facts give rise to such instances. Since $\mu^n(0) = 0$, the reflection phenomena do not arise in the original system, as well as in the

fluid and diffusion limits. Furthermore, with linear arrival and service rates, taking expectations in (3.1) provides a linear DE for $\mathbf{E}Q$. Finally, the intrinsic structure of the system at hand results in $\mathbf{E}Q^n$ and Var Q^n being linear in n [see (10.7)]. Then our rescalings (4.1) and (7.1) degenerate when applied to the corresponding mean values and variances.

State-independent networks. For such models, fluid limits are piecewise linear nonnegative functions; diffusion limits are combinations of Brownian, reflected Brownian diffusions and zero processes. Each station is permanently overloaded, critically loaded or underloaded, but with a possible initial transient phase. In this case, our results complement those of Chen and Mandelbaum [14, 15].

10.3. *Congestion-dependent dynamics in manufacturing and communication*. The queueing networks in this subsection are small-size versions of some well-known models. Our main concern is the *transient* behavior of queueing processes, while the papers from which our models originate focused on the stationary distributions of the corresponding birth and death processes. Further, we use these restricted models to explain a physical meaning of state-dependent arrival and service rates, and, especially, *state-dependent routing* policies. Our analysis is complemented by numerical examples in Section 11.1.

Flexible manufacturing systems. Examples 1 and 2 are drawn from Buzacott and Yao [12, 84, 85] and Serfozo [70].

EXAMPLE 1. An appropriate model for various flexible manufacturing systems is a queueing network with a finite population, where customers (parts) follow a *probabilistic shortest-queue routing* scheme. For example, consider the three-station network depicted in Figure 3, with the primitives

$$
\begin{aligned}
&\lambda_1(Q) = \lambda \cdot (na - Q_1 - Q_2 - Q_3)^+, \qquad \lambda_2(\cdot) = \lambda_3(\cdot) \equiv 0,\\
&\mu_k(\cdot) \equiv n\mu_k, \qquad k = 1,2,3;\\
&P(\cdot) = \begin{bmatrix} 0 & p_{12}(\cdot) & p_{13}(\cdot) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},\\
&p_{1k}(Q) = \frac{(nb_k - Q_k)^+}{n(b_2 + b_3)}, \qquad k = 2,3,
\end{aligned}
\tag{10.8}
$$

for some positive $\lambda, \mu_1, \mu_2, \mu_3, a, b_1, b_2$ and some $n \in \mathscr{Z}_+$.

REMARK. In this model, arrivals are generated by na independent sources, each of which operates at rate λ. Hence, na is the maximal number of customers in the systems (the size of the population). Such forms of $\lambda_1(\cdot)$

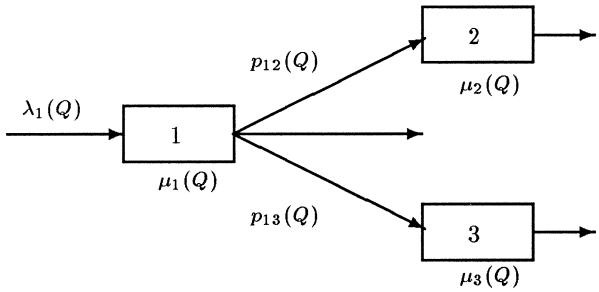


FIG. 3. *A three-station model with state-dependent routing.*

arise, for example, in repairman problems (see, e.g., [35] and [34]). Furthermore, according to (10.8), customers leaving state 1 are routed with a higher probability to the station (2 or 3) with the largest currently available waiting room. (The parameters b_2 and b_3 specify the maximal number of customers at station 2 and 3, respectively.) Note that, with probability $(1 - p_{12} - p_{13})$, customers leave the network after station 1.

It is known that finite-population models with appropriate parameters may provide reasonable approximations for closed networks. (See [79] and numerical examples in Section 11.1.) In this case, the total number of customers in the network is approximately constant. Then $(1 - p_{12} - p_{13})$ can be interpreted as the probability that a customer, after service at station 1, remains at this station due to saturation of stations 2 and 3.

Our FLLN and FCLT give fluid and diffusion limits for this network, as $n \uparrow \infty$ (that is, approximations as the population and waiting rooms grow). These limits are solutions to (4.7), (7.4) and (8.3), with

$$\lambda_1(\xi) = \lambda \cdot (a - \xi_1 - \xi_2 - \xi_3)^+, \qquad \lambda_2(\cdot) = \lambda_3(\cdot) \equiv 0;$$

$$\mu_k(\cdot) \equiv \mu_k, \; k = 1,2,3; \qquad P(\cdot) = \begin{bmatrix} 0 & p_{12}(\cdot) & p_{13}(\cdot) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$p_{1k}(\xi) = \frac{(b_k - \xi_k)^+}{b_1 + b_2}, \qquad k = 2,3; \; \xi_1, \xi_2, \xi_3 \geq 0.$$

EXAMPLE 2. Another useful model for various manufacturing systems is a *star network* with workstations linked by a material handling system and governed by reversible (probabilistic) shortest-queue routing. To be specific, consider a network consisting of K stations, with station 1 as the center (see Figure 4). Assume that $\lambda(\cdot) = n(\lambda, 0, \ldots, 0)^T$ and $\mu(\cdot) = n\mu$, for some $\lambda > 0$,

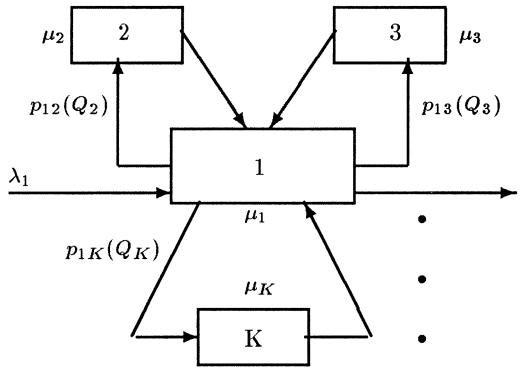


FIG. 4. *A star network with state-dependent routing*.

positive vector μ and some $n \in \mathscr{Z}_+$. Further let

$$P(\cdot) = \begin{bmatrix} 0 & p_{12}(\cdot) & \cdots & p_{1K}(\cdot) \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}, \tag{10.9}$$

$$p_{1k}(Q) = \frac{(nb_k - Q_k)^+}{n\sum_{j=1}^K b_j}, \qquad k = 2, \ldots, K,$$

for some $b_k > 0$, $k = 1, \ldots, K$, and some $n \in \mathscr{Z}_+$. Then the fluid and diffusion limits for this network are solutions to (4.7), (7.4) and (8.3), where $\lambda(\cdot) = (\lambda, 0, \ldots, 0)^T$, $\mu(\cdot) = \mu$ and $P(\cdot)$ as in (10.9), with

$$p_{1k}(\xi) = \frac{(b_k - \xi_k)^+}{\sum_{j=1}^K b_j}, \qquad k = 2, \ldots, K, \ \xi \in \mathbb{R}_+^K.$$

Computer communication networks. Models with *adaptive routing* and *adaptive rates* of processing are useful in optimization and performance evaluation of computer networks. Examples 3 and 4 below are representative of such models (see [75], [47], [10], [77] and also [12] and [84]). Example 5 is taken from [33] and [62].

EXAMPLE 3. Consider a three-station network (see Figure 3), with the primitives $\lambda(\cdot)$, $\mu(\cdot)$ and $P(\cdot)$ as in Example 1, except that

$$p_{1k}(Q) = \frac{(nb_k - Q_k)^+}{n(b_2 + b_3) - Q_2 - Q_3}, \qquad k = 2, 3; \tag{10.10}$$

for some $b_2, b_3, a > 0$.

REMARK 10.11. When the above model is used to model computer networks, the first station can be interpreted as a central processor, and the second and third stations are interpreted as peripheral devices. The meaning of parameters n, a, b_1, b_2 is similar to that in Example 1, and (10.10) describes a probabilistic shortest-queue policy. However, in contrast to Example 1, if $Q_2 + Q_3 < n(b_2 + b_3)$, then $p_{12} + p_{13} = 1$. That is, as long as b_2 and b_3 are sufficiently large (e.g., $b_2 + b_3 > a$), our model describes a network *without losses*. Reducing b_2 and b_3 introduces *losses*. We analyze numerically both of these cases in Section 11.

Fluid and diffusion limits for this network are given by (4.7), (7.4) and (8.3), with $\lambda(\cdot)$, $\mu(\cdot)$ and $P(\cdot)$ as in Example 1, except that

$$p_{1k}(\xi) = \frac{(b_k - \xi_k)^+}{b_2 + b_3 - \xi_2 - \xi_3}, \qquad k = 2, 3;\ \xi_1, \xi_2, \xi_3 \geq 0.$$

EXAMPLE 4. Consider the tandem three-station network in Figure 5, with primitives $\lambda(\cdot)$ and $\mu(\cdot)$ as in Example 1, and

$$P(\cdot) = \begin{bmatrix} p_1(\cdot) & 0 & 0 \\ 0 & p_2(\cdot) & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad p_1(Q) = \frac{Q_2 + Q_3}{nb_1}, \tag{10.12}$$

$$p_2(Q) = \frac{Q_3}{nb_2},$$

for some $b_1, b_2 > a > 0$. Here, the effective service rates at stations 1 and 2 decrease as the saturation of downstream stations increases. (The model can be easily recast as a network with state-dependent service rates and state-independent routing.)

The fluid and diffusion limits are solutions to (4.7), (7.4) and (8.3), with $\lambda(\cdot)$ and $\mu(\cdot)$ as in Example 1, $P(\cdot)$ as in (10.12) and

$$p_1(\xi) = \frac{\xi_2 + \xi_3}{b_1}, \qquad p_2(\xi) = \frac{\xi_3}{b_2}; \qquad \xi_k \geq 0,\ k = 1, 2, 3.$$

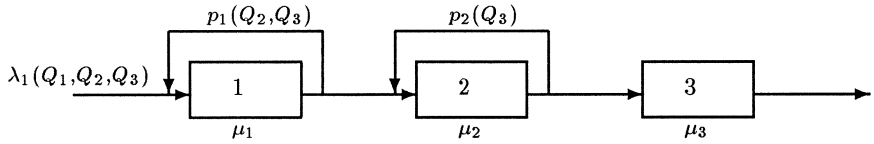


FIG. 5. *A tandem network with adaptive feedback.*

EXAMPLE 5. Another useful model is a general K-station network with state-independent arrival rates and routing, and with service rates given by

$$\mu_k(Q) = \frac{n\mu Q}{a_k^n + Q}; \qquad a_k^n = na_k, \qquad k = 1, \ldots, K, \tag{10.13}$$

for some nonnegative a_k, $k = 1, \ldots, K$, and some $n \in \mathscr{Z}_+$. In this case,

$$\mu_k(\xi) = \frac{\mu\xi}{a_k + \xi}, \qquad k = 1, \ldots, K, \ \xi \in \mathbb{R}_+^K.$$

It is notable that, in (10.13), alternatively setting $a_k^n = a_k$ or $a_k^n = \sqrt{n}\, a_k$ leads to the systems studied by Yamada ([82] and [83], respectively). (For an extended discussion on this issue, see Section 11.3 and also Sections 4.6 and 5.9 in [59]).

10.4. *Learning systems.* A manufacturing system with learning (improvement) is a system in which the time necessary to complete an operation is reduced as it is repeated over and over. The relationship that expresses this increase in service rate is called a *learning curve*. (See, e.g., [78], page 280.) We now describe a simple state-dependent network, which can be used as a basis for models of learning. Consider a two-station tandem network, as that depicted in Figure 1, but with primitives

$$\lambda_1(\cdot) \equiv n\lambda, \qquad \lambda_2(\cdot) \equiv 0; \qquad \mu_1(Q) = n\mu\left(\frac{Q_2}{n}\right),$$
$$\mu_2(\cdot) \equiv 0; \qquad P(\cdot) = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

for some $n \in \mathscr{Z}_+$, $\lambda > 0$ and an increasing positive function μ. The first station is a system with learning, and μ is the corresponding learning curve. In practice, the learning curve is typically of the form $\mu(\xi) = c(1 + \xi)^\alpha$, $c > 0$, $\alpha \in (0, 1)$.

The model presented above animates the one in [78] as a state-dependent networks. (In [78], learning systems are characterized by the forward equations for the corresponding birth and death processes.)

The fluid and diffusion limits arise as $n \uparrow \infty$, that is, when arrival and service rates become large. These limits are solutions to (4.7), (7.4) and (8.3), with

$$\lambda_1(\cdot) \equiv \lambda, \qquad \lambda_2(\cdot) \equiv 0; \qquad \mu_1(\xi) = \mu(\xi_2), \qquad \mu_2(\cdot) \equiv 0;$$
$$P(\cdot) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \xi_1, \xi_2 \geq 0.$$

Some practically interesting problems that arise within the context of learning models are listed:

1. Investigating the stabilizing effect of learning on systems subject to large constant, increasing, periodic or other *time-inhomogeneous* arrival process (representing, for example, an unexpected surge of demands.)

2. Studying multiserver systems governed by a *machine-release* policy [78]: machines are released from the system when, due to learning, processing is fast enough so that utilization falls below some predetermined threshold.

Such examples will be presented in [56], devoted to time-dependent networks. (See also Section 15, where we relate *time-* and *state-dependent* queueing networks.)

10.5. *Epidemic models.* Another interesting field of applications is epidemics, namely, spreads of infections. A simple model of a stochastic epidemic can be described as follows (see [36]). A population is subdivided into three classes (groups): those who are susceptible to the infection, those who are infected and those who have recovered and are immune to reinfection. The class of infectives is further subdivided into subclasses according to stages of the incubation period and the progress of disease. Infections occur at a rate proportional to the current number of individuals in the classes of both susceptibles and infectives. This model of stochastic epidemic can be represented as a K-station tandem network (see Figure 6) with the primitives

$$\begin{aligned}
&\lambda_1(\cdot) = \cdots = \lambda_K(\cdot) = \mu_K(\cdot) \equiv 0,\\
&\mu_1(Q) = aQ_1 \cdot \frac{b_2Q_2 + \cdots + b_{K-1}Q_{K-1}}{n},\\
&\mu_k(Q) = c_kQ_k, \qquad k = 2,\ldots,K-1,\\
&p_{jk}(Q) = \begin{cases} 1, & j = 1,\ldots,K-1,\ k = j+1,\\ 0, & \text{otherwise,}\end{cases}
\end{aligned}$$

for some positive a, b_k, c_k, $k = 1,\ldots,K$, and some $n \in \mathscr{Z}_+$. According to the description above, Q_1 is the current number of individuals susceptible to the infection, $Q_2,\ldots,Q_{K-1}$ are the numbers of individuals in the $K-2$ subclasses of the group of infectives and Q_K is the number immune to reinfection.

The fluid and diffusion limits for this network are solutions to (4.7), (7.4) and (8.3), with

$$\begin{aligned}
&\lambda_1(\cdot) = \cdots = \lambda_K(\cdot) = \mu_K(\cdot) \equiv 0,\\
&\mu_1(\xi) = a\xi_1 \cdot (b_2\xi_2 + \cdots + b_{K-1}\xi_{K-1}),\\
&\mu_k(\xi) = c_k\xi_k, \qquad k = 2,\ldots,K-1,\\
&p_{jk}(\xi) = \begin{cases} 1, & j = 1,\ldots,K-1,\ k = j+1;\\ 0, & \text{otherwise,}\end{cases} \qquad \xi \in \mathbb{R}_+^K.
\end{aligned}$$

FIG. 6. *A queueing model of epidemics.*

Note that we can easily incorporate various sources of heterogeneity between individuals by adding a new station for each homogeneous group.

Isham [36] analyzed several stochastic epidemic models. In particular, using Gaussian diffusion approximations, she estimated the first two moments of processes of interest. Our approximations lead exactly to the same results. (We perform similar calculations, applied to other models, in Section 11.) The theoretical justification for [36] is the FCLTs presented, for example, in [7], [48] and [49]. Note that, in epidemic models, limits do not involve the reflection phenomenon, and therefore the corresponding limit theorems are a particular case of our FLLN and FCLT.

10.6. *Data networks with bursty sources.* The model we now describe was studied from various viewpoints by Anick, Mitra and Sondhi [3], Knessl and Morrison [43] and Kushner and Martins [52].

Consider a data-transmission system, which receives messages from n independent sources. The sources alternate between *on* and *off* states and create messages during the *on* periods. Assume that the duration of each *on* (*off*) period is exponentially distributed with mean value $1/\eta$ $(1/\nu)$. During an *on* period, each source sends messages according to a Poisson process with rate λ. The service time of the transmission system is exponentially distributed with mean value $1/n\mu$. This model can be represented as a queueing network that consists of two nonlinked stations, with parameters

$$\lambda_1(Q) = \lambda \cdot Q_2, \qquad \lambda_2(Q) = \nu \cdot (n - Q_2)^+; \qquad \mu_1(\cdot) \equiv n\mu,$$
$$\mu_2(Q) = \eta \cdot (Q_2 \wedge n); \qquad P(\cdot) \equiv 0.$$

The first station represents the data-transmission system, while the second station is introduced to model *on* and *off* periods of the sources: being at station 2 corresponds to being *off*.

The fluid and diffusion limits arise as $n \uparrow \infty$; that is, as the number of sources and the rate of service become large. The limits are solutions to (4.7), (7.4) and (8.3), with

$$\lambda_1(\xi) = \lambda \cdot \xi_2, \qquad \lambda_2(\xi) = \nu \cdot (1 - \xi_2)^+; \qquad \mu_1(\cdot) \equiv \mu,$$
$$\mu_2(\xi) = \eta \cdot (\xi_2 \wedge 1); \qquad P(\cdot) \equiv 0; \qquad \xi_1, \xi_2 \geq 0.$$

For example, if $q(0) = [0, \nu/(\nu + \eta)]^T$ and $\mu = \lambda\nu/(\nu + \eta)$, then we get the expressions in [52]:

$$q \equiv q(0),$$
$$dV_1(t) = \lambda V_2(t)\, dt + \sqrt{2\frac{\lambda\nu}{\nu + \eta}}\, dW_1(t) + dY(t),$$
$$dV_2(t) = -(\nu + \eta)V_2(t)\, dt + \sqrt{2\frac{\eta\nu}{\nu + \eta}}\, dW_2(t).$$

10.7. *Multiprocessor systems with breakdowns.* By analogy with Section 10.6, we can construct a multiserver (multiprocessor) system, where each of

the servers is subject to independent random breakdowns and repairs. Such a system was considered by Mitrani and Puhalskii [63].

To be specific, assume that there are n identical independent parallel processors. The processors alternate between *on* and *off* and they are operative during the *on* periods. Further, suppose that the duration of each *on* (respectively, *off*) period is exponentially distributed with mean value $1/\eta$ (respectively $1/\nu$). Jobs arrive to the system according to a Poisson process with the rate $n\lambda$. The service time is exponentially distributed with mean value $1/\mu$. This model can be represented as a queueing network that consists of two nonlinked stations and has the following parameters:

$$\lambda_1(\cdot) \equiv n\lambda, \qquad \lambda_2(Q) = \nu \cdot (n - Q_2)^+, \qquad \mu_1(Q) = \mu \cdot (Q_1 \wedge Q_2),$$
$$\mu_2(Q) = \eta \cdot (Q_2 \wedge n); \qquad P(\cdot) \equiv 0.$$

As in Section 10.6, the first station is actually the multiprocessor system, while the second station models *on* and *off* periods of the processors. This model animates the birth and death processes from [63] as a state-dependent queueing network.

The fluid and diffusion limits arise as $n \uparrow \infty$; that is, as the number of processors and the rate of service become large. These limits are solutions to (4.7), (7.4) and (8.3), with

$$\lambda_1(\xi) = \lambda, \qquad \lambda_2(\xi) = \nu \cdot (1 - \xi_2)^+; \qquad \mu_1(\xi) = \mu \cdot (\xi_1 \wedge \xi_2),$$
$$\mu_2(\xi) = \eta \cdot (\xi_2 \wedge 1); \qquad P(\cdot) \equiv 0; \qquad \xi_1, \xi_2 \geq 0.$$

REMARK. Setting $\mu_2(Q) = \eta \cdot (Q_2 \wedge cn)$, for some $c \in (0, 1]$, leads to a more general system, in which processors may be forced to wait in queue for repair.

10.8. *Multiserver systems with breakdowns and blocking*. Consider a two-station tandem system. Each station is a multiprocessor system with breakdowns of processors (as in Section 10.7). A distinguishing feature of this system is that the buffer at the second station has a finite capacity. The service rate of the first station is adapted to the buffer content of the second by having fewer servers work when buffer content is high. In particular, when the buffer is full, all the servers at the first station stop serving. [See $\mu_1(\cdot)$ below.] An appropriate model for this system is a four-station queueing network with the primitives

$$\lambda_1(\cdot) \equiv n\lambda, \qquad \lambda_2(\cdot) \equiv 0, \qquad \lambda_3(Q) = \nu_3 \cdot (a_3 n - Q_3)^+,$$
$$\lambda_4(Q) = \nu_4 \cdot (a_4 n - Q_4)^+;$$
$$\mu_1(Q) = \mu \cdot \left[Q_1 \wedge (bn - Q_2)^+ \wedge Q_3\right], \qquad \mu_2(Q) = \hat{\mu} \cdot [Q_2 \wedge Q_4],$$
$$\mu_3(Q) = \eta_3 \cdot (Q_3 \wedge c_3 n), \qquad \mu_4(Q) = \eta_4 \cdot (Q_4 \wedge c_4 n);$$
$$p_{jk}(Q) = \begin{cases} 1, & j = 1, k = 2, \\ 0, & \text{otherwise}, \end{cases}$$

for some positive $\lambda, \mu, \hat{\mu}, b, \nu_k, \eta_k, c_k$, $k = 3,4$, and some $n \in \mathcal{Z}_+$. In this model, stations 1 and 2 are the multiprocessor systems. Stations 3 and 4 model breakdowns of processors at stations 1 and 2, respectively.

The fluid and diffusion limits for this network are solutions to (4.7), (7.4) and (8.3), with

$$\lambda_1(\cdot) \equiv \lambda, \qquad \lambda_2(\cdot) \equiv 0, \qquad \lambda_3(\xi) = \nu_3 \cdot (a_3 - \xi_3)^+,$$
$$\lambda_4(\xi) = \nu_4 \cdot (a_4 - \xi_4)^+;$$
$$\mu_1(\xi) = \mu \cdot \left[\xi_1 \wedge (b - \xi_2)^+ \wedge \xi_3\right], \qquad \mu_2(\xi) = \hat{\mu} \cdot [\xi_2 \wedge \xi_4],$$
$$\mu_3(\xi) = \eta_3 \cdot (\xi_3 \wedge c_3), \qquad \mu_4(\xi) = \eta_4 \cdot (\xi_4 \wedge c_4);$$
$$p_{jk}(\xi) = \begin{cases} 1, & j = 1, k = 2, \\ 0, & \text{otherwise}, \end{cases} \qquad \xi_k \geq 0,\ k = 1, \ldots, 4.$$

The system described above is analogous to that considered by Mitra [61], but our blocking mechanism is different. Specifically, we change μ_1 gradually with Q_2, while in [61], rates change in an abrupt fashion only when buffers are either full or empty. We can incorporate the latter blocking mechanism in our state-dependent framework by using piecewise linear service rates, similar to those used in Section 11.2, to model finite buffers.

10.9. *Stochastic traffic assignment models.* We can show that various stochastic traffic assignment models, as in Davis and Nihan [20], can be modeled by a queueing network with state-dependent routing probabilities

$$p_{jk}(Q) = \frac{\exp\left(-c_{jk}(Q)\,\alpha\right)}{\sum_{i=1}^{K} \exp\left(-c_{ji}(Q)\,\alpha\right)};$$
$$c_{jk}(Q) = \frac{\beta_{jk}^1 Q_1 + \cdots + \beta_{jk}^K Q_K}{n}, \qquad j, k = 1, \ldots, K,$$

for some nonnegative α, β_{jk}^i, $j, k, i = 1, \ldots, K$, and some $n \in \mathcal{Z}_+$. It is shown in [20] that, as the number of individual travelers becomes large, the network's traffic volumes can be approximated by the sum of a nonlinear deterministic function and a time-varying linear Gaussian process. These approximations correspond to our fluid and diffusion limits.

10.10. *Human-service systems.* Many queues that are encountered in our life are *state-* and *time-dependent*. Some examples are public service centers, telephone systems, banks, hospitals and others. In these systems, customers react to state changes: they typically prefer short queues, jockey, renege and so on. State-dependent queueing networks provide, therefore, a natural framework for design, performance analysis and optimization of service systems. An example of using state-dependent queues for approximate analysis of service was given by Worthington [81], who applied models of queues with reneging (see Section 5.8 in [59]) to the hospital waiting-list problem.

Other examples are [80] and [39] (see also [76]). They considered the problem of finding the right number of servers in multiserver service systems, so as to keep the probability of delay under some predetermined level. In [80], the staffing problem was solved by infinite-server approximations for a system with a *time-homogeneous* arrival process. In contrast, the arrival process in [39] is *time-inhomogeneous*; hence, the "right" number of servers s becomes time-dependent as well. In other words, when there are $s(t)$ servers in the system, the service rate is $\mu(Q,t) = \mu \cdot (Q \wedge s(t))$ and the problem thus is to choose the right function $s(t)$. Actually, such a system is representative of queues, which are both state- and time-dependent. Approximating such systems is important for future research (see also Section 15).

11. Numerical examples. The numerical examples in this section constitute an attempt to demonstrate the quality of our fluid and diffusion approximations and to demonstrate their use in facilitating the analysis of various queueing networks. Section 11.1 is devoted to a three-station network (see Figure 3) governed by various routing policies. In Section 11.2, we show that our approximations also fit systems with large *finite buffers* despite the fact that they are derived for networks with *infinite buffers*. In Section 11.3, we compare different rescaling procedures, by applying them to a multiserver queue.

Consider a sequence $(M_\xi^n/M_\xi^n/1)^K$, $n = 1, 2, \ldots$, of state-dependent networks, which satisfies the conditions of FLLN and FCLT (Theorems 4.6 and 7.2 and Proposition 8.2). The theorems suggest that, for sufficiently large n,

$$Q^n(\cdot) \overset{d}{\sim} nq(\cdot) + \sqrt{n}\, V(\cdot), \tag{11.1}$$

$$\mathbf{E}Q^n(\cdot) \sim nq(\cdot) + \sqrt{n}\, \mathbf{E}V(\cdot), \qquad \operatorname{Cov} Q^n(\cdot) \sim n \operatorname{Cov} V(\cdot), \tag{11.2}$$

$$Y^n(\cdot) \overset{d}{\sim} ny(\cdot) + \sqrt{n}\, H(\cdot), \tag{11.3}$$

$$\mathbf{E}Y^n(\cdot) \sim ny(\cdot) + \sqrt{n}\, \mathbf{E}H(\cdot), \qquad \operatorname{Cov} Y^n(\cdot) \sim n \operatorname{Cov} H(\cdot). \tag{11.4}$$

These relations justify our methods below for approximating queueing and idle-time processes by their corresponding fluid and diffusion limits.

REMARK. Equations (11.1) and (11.2) suggest, at least formally, that also

$$\begin{aligned} Q^n(\infty) &\overset{d}{\sim} nq(\infty) + \sqrt{n}\, V(\infty), \\ \mathbf{E}Q^n(\infty) &\sim nq(\infty) + \sqrt{n}\, \mathbf{E}V(\infty), \\ \operatorname{Cov} Q^n(\infty) &\sim n \operatorname{Cov} V(\infty), \end{aligned} \tag{11.5}$$

assuming, of course, that the corresponding stationary values and distributions exist. A rigorous justification of (11.5) is not available to the best of our knowledge. Examples of theorems that support such approximations are given in [29], [41], [25], Chapter 4, Section 9, and [53]. Since our focus is on the transient behavior of networks, we do not pursue this further here.

Our analysis combines the following tools:

1. Analytical solution of DEs (fluid) and SDEs (diffusion) whenever possible.
2. Numerical solution of DEs and SDEs (as in [42]) using MATLAB or customized software.
3. Simulation of the original queueing systems using SIMAN/ARENA [66].

11.1. *Networks with state-dependent routing*. This subsection is devoted to analysis of the three-station network depicted in Figure 3 and described by Examples 1 and 3 in Section 10.3. By means of fluid approximations, we compare different routing strategies given by (10.8), (10.10) and others. As pointed out in Section 10.3, this model captures significant features of many manufacturing and computer systems (see [47], [75] and [84]).

In [47], [75] and [84], the focus is on the stationary phase and numerical results pertain to a small number of customers in the network. In contrast, our goal here is to analyze large systems in their transient phase. In addition, we attempt to show the following results:

1. Our fluid limits provide reasonable approximations for queues in general and for idle times of overloaded stations.
2. Our fluid approximations are useful for comparing different routing policies and aid in the identification of close-to-optimal modes of operation.
3. The state-dependent routing (10.10) can be used to approximate the shortest-queue routing policy. (The analysis of the latter is often intractable.) Moreover, using our state-dependent routing leads to improved performance of the network.
4. Our open-network models can approximate closed networks.

Through our numerical experiments, we seek to improve or optimize a set of performance measures. We now describe these measures, which arise from interpretation of the network as a computer or manufacturing systems (see Section 10.3 and the references cited above):

Performance criteria.

Throughput: The potential of the two peripheral devices (stations 2 and 3) should be fully realized (they should be critically loaded or overloaded). Then the system throughput is the total service rate at stations 2 and 3.

Queues: The magnitude of the queue at the central processor should be relatively small (station 1 should be underloaded or at most critically loaded).

Balance: The operation of stations 2 and 3 should be balanced, in the sense of similar magnitudes of queues (even though service rates may differ). The queues at stations 2 and 3 should be bounded.

Blocking: If stations 2 and 3 have limited buffer capacities, then the probability of blocking should be low. (In the model considered, customers blocked at station 1 leave the network and are considered lost.)

Stability: If there exists a stationary distribution, then the transient phase should be relatively short.

As a start, consider the three-station network, of Example 3 in Section 10.3. Recall from Remark 10.11 that as long as b_2 and b_3 are sufficiently large, our model describes a network without losses. In contrast, reducing b_2 and b_3 introduces losses. We consider these cases in turn below.

Networks without losses. *Comparison of fluid approximations with simulation*: In Figure 7, we compare the queueing and idle-time processes computed from 300 simulations and from numerical solution of DEs for fluid approximations. The following parameters were chosen:

$$\begin{aligned} &\lambda = 5, \quad a = 10, \quad \mu_1 = 10, \quad \mu_2 = 2, \\ &\mu_3 = 7, \quad b_2 = b_3 = 5, \quad Q(0) = 0. \end{aligned} \tag{11.6}$$

Figure 7a and b exhibits data for $\mathbf{E}Q^n$ and $\mathbf{E}I^n$ for $n = 100$; Figure 7c and d exhibits data for $\mathbf{E}Q^n$ and $\mathbf{E}I^n$ for $n = 1000$. A comparison between Figure 7a and b and c and d demonstrates that the quality of the fluid approximation improves as n increases.

REMARK. Approximations for the idle-time processes by the fluid limits are less satisfactory than those for the queueing processes. We can improve these approximations through the second-order diffusion refinement in Proposition 8.2 [see (11.3) and (11.4)].

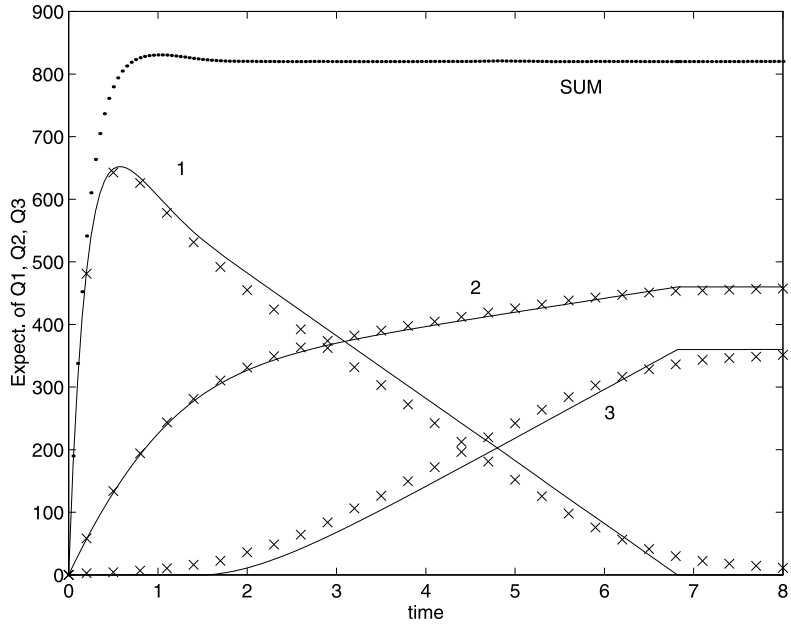
The fluid limit can produce even better approximations. Figure 8, demonstrates this for the network with $\lambda, a, \mu_1, \mu_2, b_2, b_3$ and $Q(0)$ taken as in (11.6), $\mu_3 = 2$ and $n = 100$, where

$$\begin{aligned} &\lambda = 5, \quad a = 10, \quad \mu_1 = 10, \quad \mu_2 = \mu_3 = 2, \\ &b_2 = b_3 = 5, \quad Q(0) = 0, \quad n = 100. \end{aligned} \tag{11.7}$$

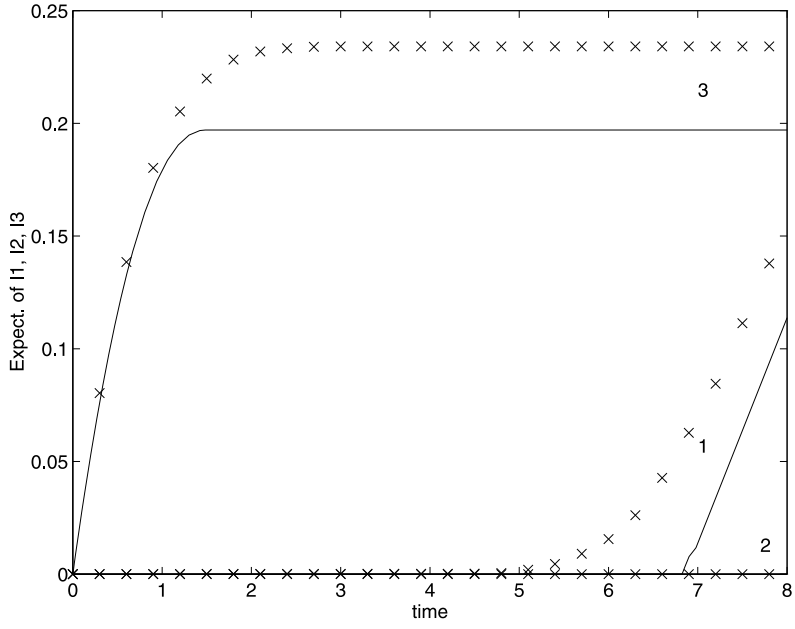
The fit here is almost perfect. Based on empirical experience, we attribute this to the symmetry $\mu_2 = \mu_3$. In the sequel, we focus on asymmetric cases, when $\mu_1 > \mu_2 \vee \mu_3$ and $\mu_2 \neq \mu_3$ [as in (11.6)]. It is explained in [75] that such a combination of parameters is the most unfavorable, from the viewpoint of the performance criteria described above. Hence, the advantages of our state-dependent routing are the most pronounced. In particular, a comparison of Figure 7 with Figure 8 shows that in the latter symmetric network, the operation is the same as that of a network with state-independent routing $p_{12} = p_{13} = 1/2$, and the transient phase is relatively short.

Analysis of Figure 7a and b leads to the following observations:

1. Station 1 is overloaded until $t \approx 6.8$ and underloaded thereafter; station 2 is permanently overloaded; station 3 is underloaded until $t \approx 1.5$ and overloaded thereafter.
2. At $t \approx 6.8$, the network enters the *stationary phase*, in the sense that the fluid approximation remains constant thereafter. The evolution during the *transient phase* cannot be deduced from exact analysis. The following calculations provide insight into the stationary behavior. Within the sta-

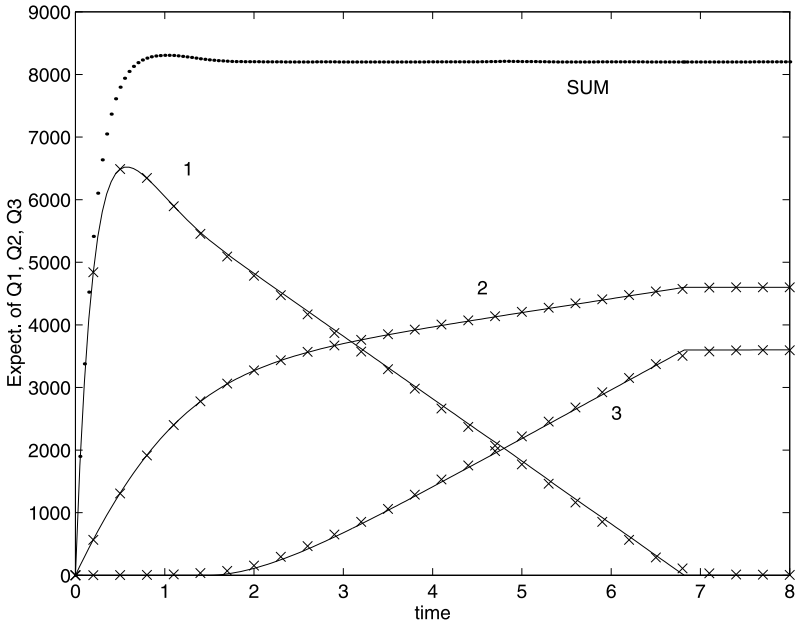


(a)

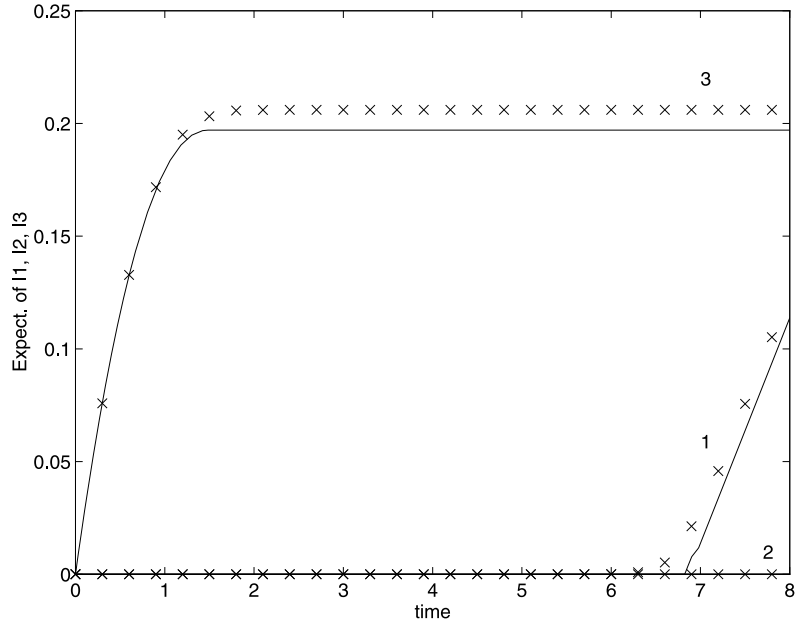


(b)

FIG. 7. *Fluid approximations versus simulation for the three-station network without losses*; *parameters are as in* (11.6). *The solid lines are computed from fluid approximations. The* $\times$*-lines are computed from simulations*: (a) $\mathbf{E}Q_1^n$, $\mathbf{E}Q_2^n$, $\mathbf{E}Q_3^n$, $\mathbf{E}(Q_1^n + Q_2^n + Q_3^n)$; $n = 100$; (b) $\mathbf{E}I_1^n$, $\mathbf{E}I_2^n$, $\mathbf{E}I_3^n$; $n = 100$; (c) $\mathbf{E}Q_1^n$, $\mathbf{E}Q_2^n$, $\mathbf{E}Q_3^n$, $\mathbf{E}(Q_1^n + Q_2^n + Q_3^n)$; $n = 1000$; (d) $\mathbf{E}I_1^n$, $\mathbf{E}I_2^n$, $\mathbf{E}I_3^n$; $n = 1000$.



(c)



(d)

FIG. 7. *Continued.*

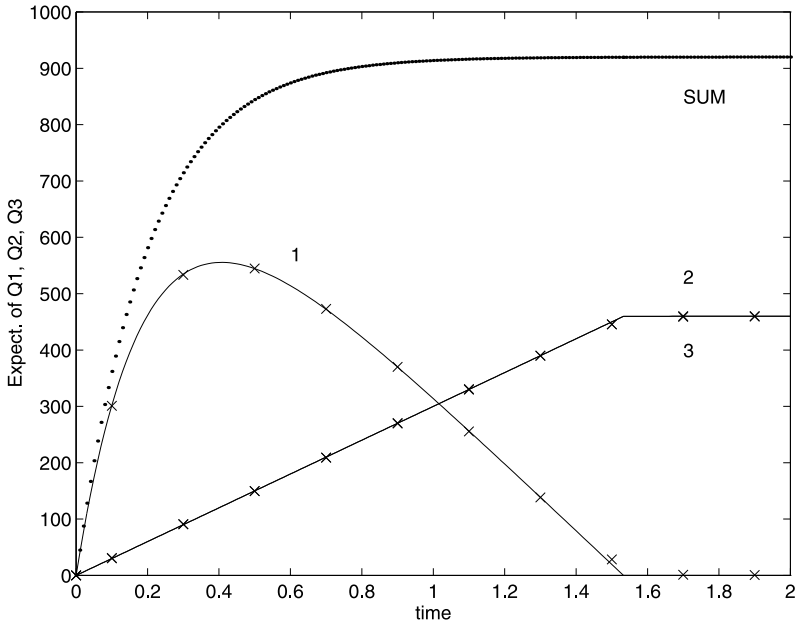


FIG. 8. *Fluid approximations versus simulation for the three-station network without losses*; *the parameters are given by* (11.7). *The solid lines are computed from fluid approximations*. *The* $\times$*-lines are computed from* 300 *simulations*. (a) $\mathbf{E}Q_1^n$, $\mathbf{E}Q_2^n$, $\mathbf{E}Q_3^n$, $\mathbf{E}(Q_1^n + Q_2^n + Q_3^n)$; (b) $\mathbf{E}I_1^n$, $\mathbf{E}I_2^n$, $\mathbf{E}I_3^n$.

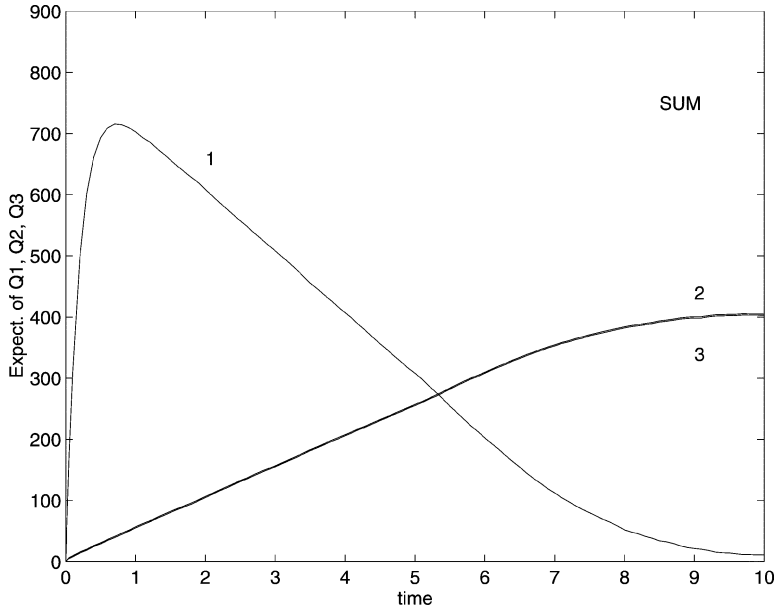
tionary phase, stations 2 and 3 are overloaded and, hence, the throughput of the network is $\mu_2 + \mu_3 = 900$. Since $\lambda_1 = 5 \cdot (1000 - Q_1 - Q_2 - Q_3)^+$, we have that during this phase $Q_1 + Q_2 + Q_3 = 820$. Furthermore, the stationary value of the traffic intensity at station 1 is $\rho = 900/1000$ and, hence, the stationary value of Q_1 is $\rho/(1 - \rho) = 9$. (This small queue corresponds to zero fluid approximation.) Next, the fluid approximations demonstrate that the stationary values of queues at stations 2 and 3 are 460 and 360, respectively. Therefore, by (10.10), we have that $p_{12} = 2/9$ and $p_{13} = 1/9$ within the stationary phase; that is, $p_{12}/p_{13} = \mu_2/\mu_3$. This relation means that the faster service is loaded more. We analyze below this routing policy and show that it is less effective than the state-dependent routing (10.10).

Comparison of different routing policies: Figure 9 compares four different routing policies: (a) shortest-queue routing; (b) state-dependent routing, given by (10.10) with $b_2 = b_3 = 4.5$; (c) state-independent routing with $p_{12} = p_{13} = 0.5$; (d) state-independent routing with $p_{12}/p_{13} = \mu_2/\mu_3$ (load the faster server more). The other parameters are chosen as in (11.6). Figure 9 yields the following observations:

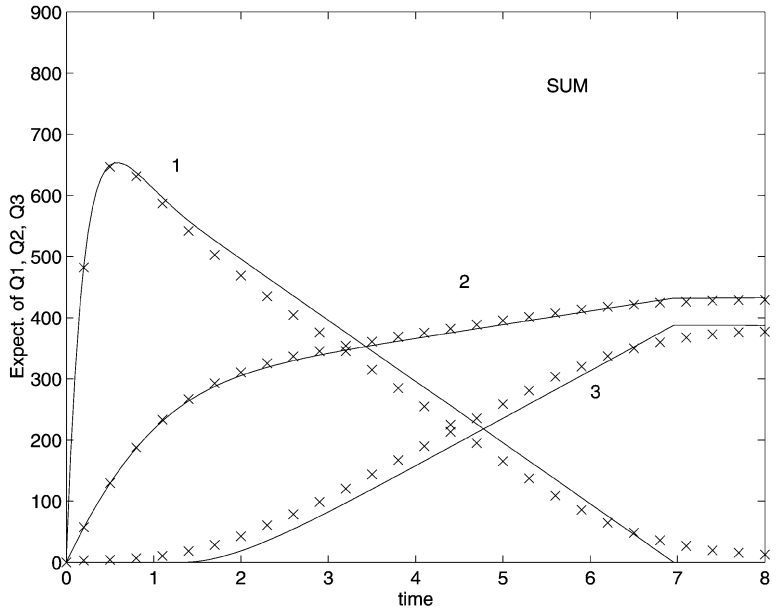
1. Policy (a) has the best performance measures: (i) The potential of stations 2 and 3 is fully realized. The stationary throughput is equal to 200 + 700 = 900. (ii) Operation of stations 2 and 3 is balanced in the sense of equal queues, despite different service rates. The queues at these stations do not exceed 400. (iii) The queue at station 1 is asymptotically (for large t) small. (Station 1 is asymptotically critically loaded.)
2. Policy (b) achieves performance that is close to that of policy (a). However, our state-dependent routing is theoretically more tractable than the shortest-queue policy.
3. Under policy (c), the operation has the worst performances. The stationary throughput of the system is as low as 410. [The stationary value of $Q_1 + Q_2 + Q_3$ is approximately 916; hence, $\lambda_1 = 5 \cdot (1000 - 916) = 420$. The stationary throughput equals, therefore, $200 + 420/2 = 410$.] The value of Q_2 is very high, while stations 1 and 3 are underloaded.
4. Policy (d) is better than policy (c), but worse than policy (b).

Networks with losses. Reducing b_2, b_3 (i.e., reducing the permissible queues at stations 2 and 3) introduces losses into the system. We compare performances under the following policies: (a) state-dependent routing, given by (10.10) with $b_2 = b_3 = 4$; (b) finite buffer system with buffers $b_2 = b_3 = 4$ and $p_{12} = p_{13} = 0.5$; (c) finite buffer system, with buffers $b_2 = b_3 = 4$ and $p_{12}/p_{13} = \mu_2/\mu_3$; (d) state-dependent routing given by (10.8) with $b_2 = b_3 = 4$. We chose the same parameters as in (11.6), except that $b_2 = b_3 = 4$, and we take $n = 100$. Figure 10 yields the following conclusions:

1. Policy (a) leads to the best performance: (i) The potential of stations 2 and 3 is fully utilized. The stationary throughput is equal to 200 + 700 = 900.

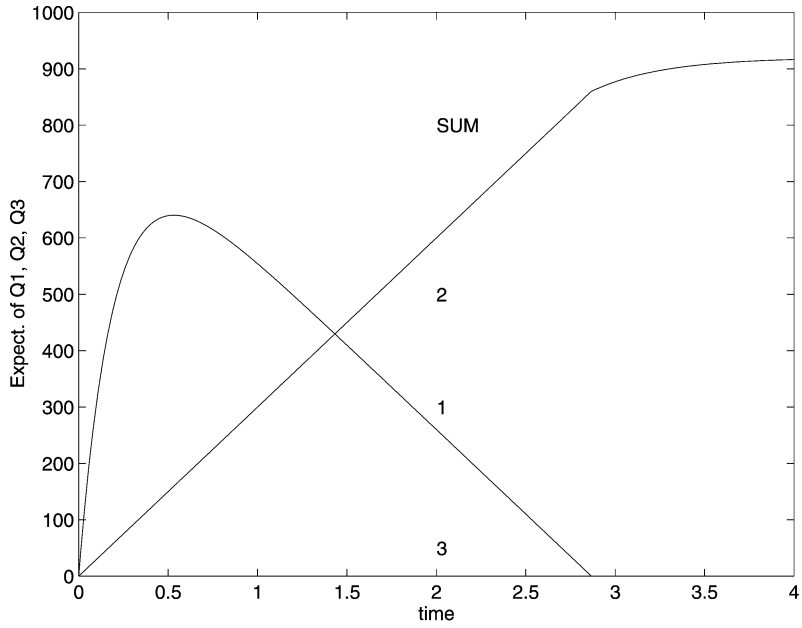


(a)

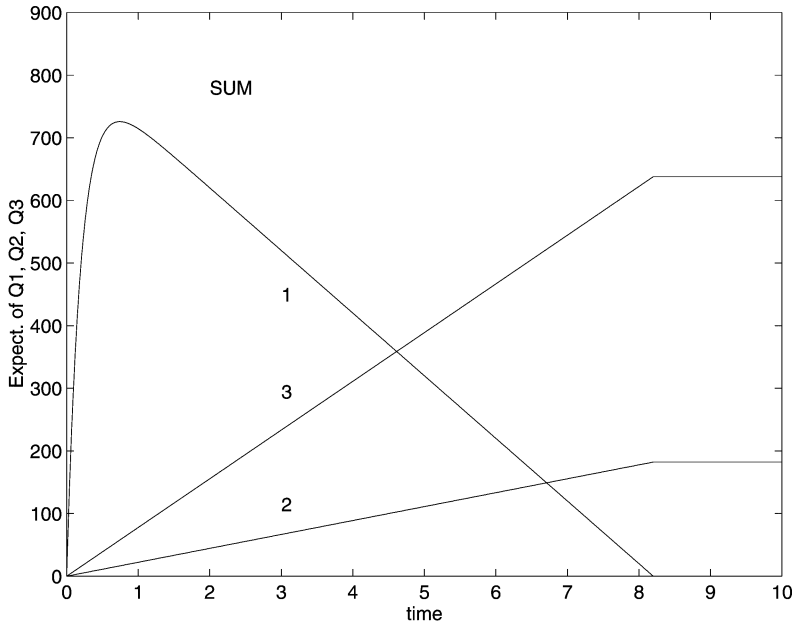


(b)

FIG. 9. *Comparing routing policies for the three-station model without losses*: $\mathbf{E}Q_1^n$, $\mathbf{E}Q_2^n$, $\mathbf{E}Q_3^n$, $\mathbf{E}(Q_1^n + Q_2^n + Q_3^n)$; $n = 100$: (a) *Shortest queue routing* (*from simulations*). (b) *State-dependent routing*, *given by* (10.10) *with* $b_2 = b_3 = 4.5$. *The solid lines are computed from fluid approximations. The* $\times$*-lines are computed from simulations.* (c) *State-independent routing with* $p_{12} = p_{13} = 0.5$; *fluid approximations.* (d) *State-independent routing with* $p_{12}/p_{13} = \mu_2/\mu_3$; *fluid approximations.*

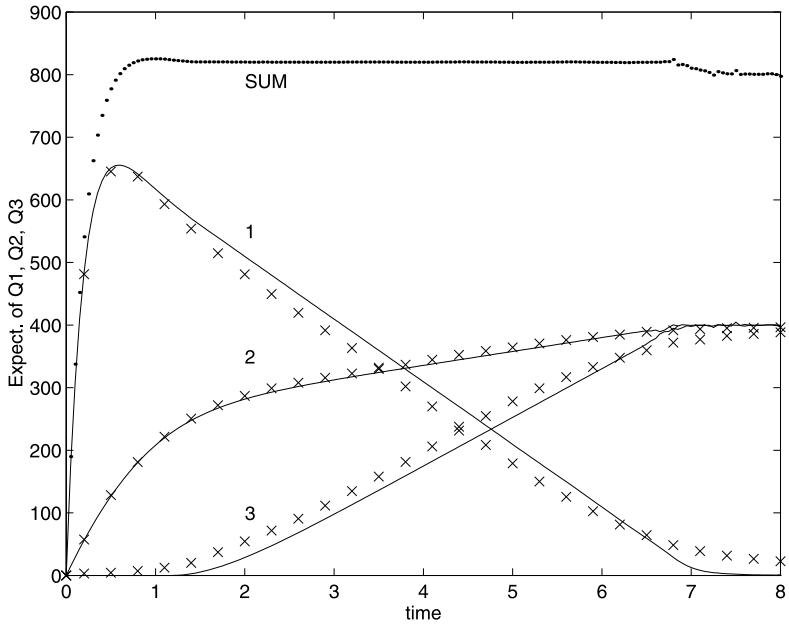


(c)

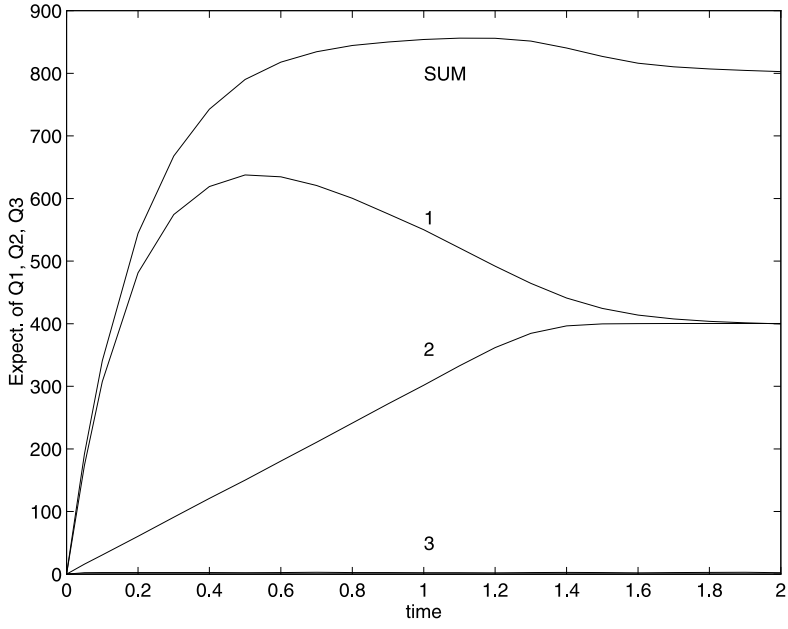


(d)

FIG. 9. *Continued.*

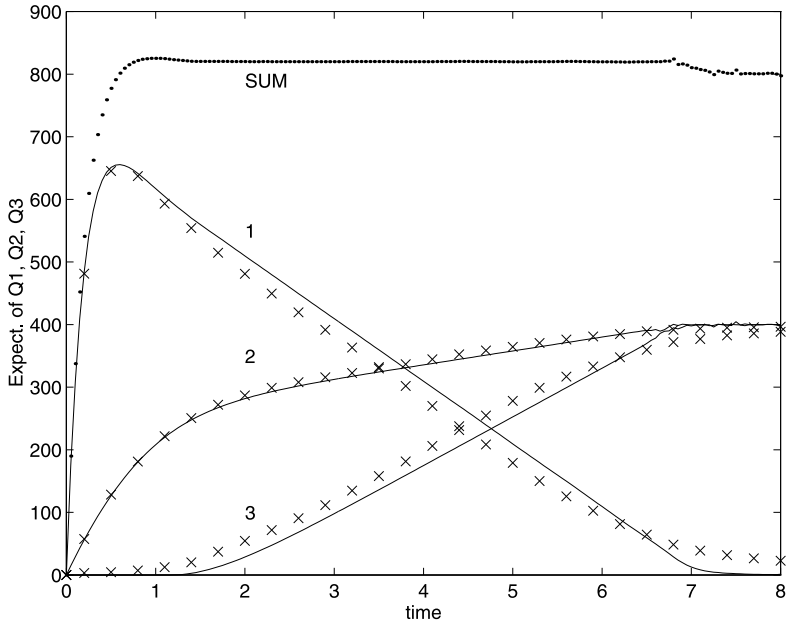


(a)

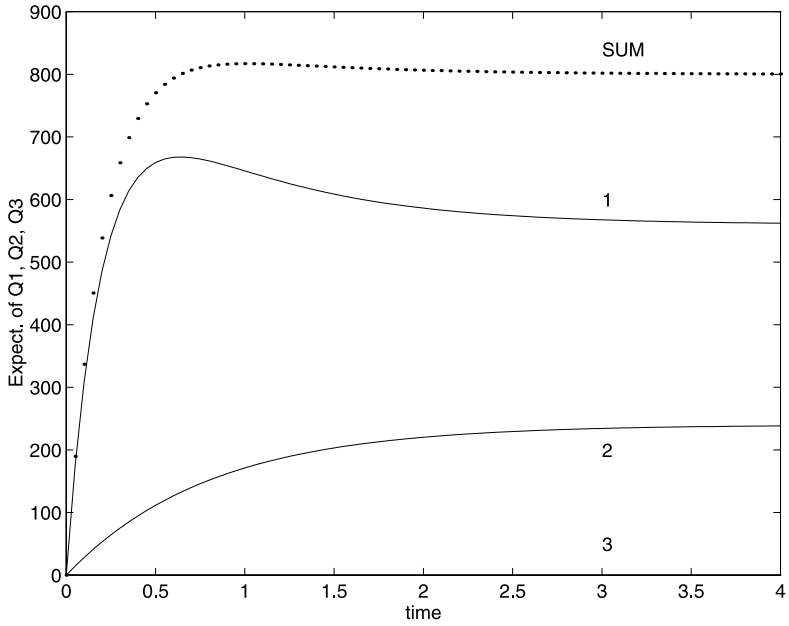


(b)

FIG. 10. *Comparison of routing policies for the three-station network with losses*: $\mathbf{E}Q_1^n$, $\mathbf{E}Q_2^n$, $\mathbf{E}Q_3^n$, $\mathbf{E}(Q_1^n + Q_2^n + Q_3^n)$; $n = 100$. (a) *State-dependent routing given by* (10.10) *with* $b_2 = b_3 = 4$. (b) *Systems with finite buffers* $b_2 = b_3$ *and state-independent routing* $p_{12} = p_{13} = 0.5$ (*simulation*). (c) *Systems with finite buffers* $b_2 = b_3$ *and state-independent routing with* $p_{12}/p_{13} = \mu_2/\mu_3$ (*simulation*). (d) *State-dependent routing given by* (10.8) *with* $b_2 = b_3 = 4$.



(c)



(d)

FIG. 10. *Continued.*

(ii) Operation of stations 2 and 3 is balanced in the sense of *asymptotically* equal queues, despite different service rates. The queues at these stations do not exceed 400. (iii) The queue at station 1 is asymptotically small. (This station is asymptotically critically loaded.) (iv) The stationary probability of losses is as small as 0.08.

2. Other policies are inferior to policy (a). In particular: (i) The stationary throughput is $200 + 1000/2 = 700$ under policies (b) and (d), while the first station is overloaded. (ii) Under policy (c) the throughput is 900, but the first station is overloaded. (iii) The stationary probability of losses approximately equals 0.3.

REMARK. Note that the total number of customers in the networks in both Figures 9 and 10 stabilizes fast into a constant. This suggests that our approximations for open networks can be used to approximate corresponding closed systems. The issue was addressed in detail by Whitt [79].

11.2. *Large finite buffers.* The networks considered in this paper have unlimited buffers. However, our results can be applied to approximate networks with large buffers (of order n). Such models arise, for example, in large human-service systems, communication networks and others, where waiting rooms are made sufficiently large to assure that blocking rarely occurs.

We approximate a single station $M/M/(C+1)$ whose arrival rate is λ, service rate μ and buffer size C, where

$$\lambda = 2n, \qquad \mu = n, \qquad C = n \tag{11.8}$$

for $n = 1000$. The approximation is a *state-dependent* single station with an *infinite buffer* and parameters

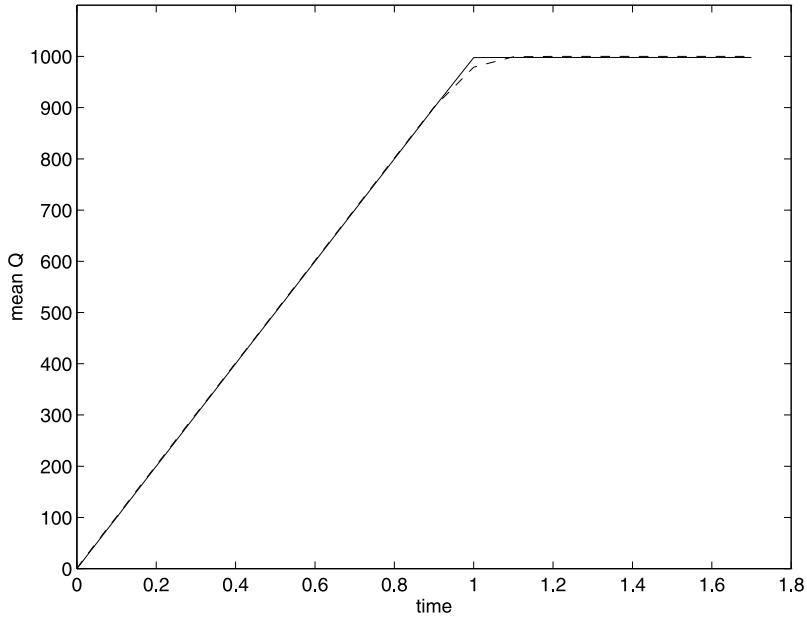
$$\lambda(Q) = 2 \cdot \left\{ n - \frac{[Q - (C - n\varepsilon)]^+}{\varepsilon} \right\}^+, \qquad \mu = n;\ C = n, \tag{11.9}$$

for $n = 1000$. We assert that, with an appropriate choice of $\varepsilon \ll C/n$, the fluid and diffusion limits for this model provide reasonable approximations for the original system with finite buffer. The rationale for this approximation is that $\lambda(Q)$ is constant up to $Q = C - n\varepsilon$ and it vanishes for $Q \geq C$. Thus, when $\varepsilon \ll C/n$, $\lambda(Q)$ is close to the rate of effective arrivals in the finite-buffer model.

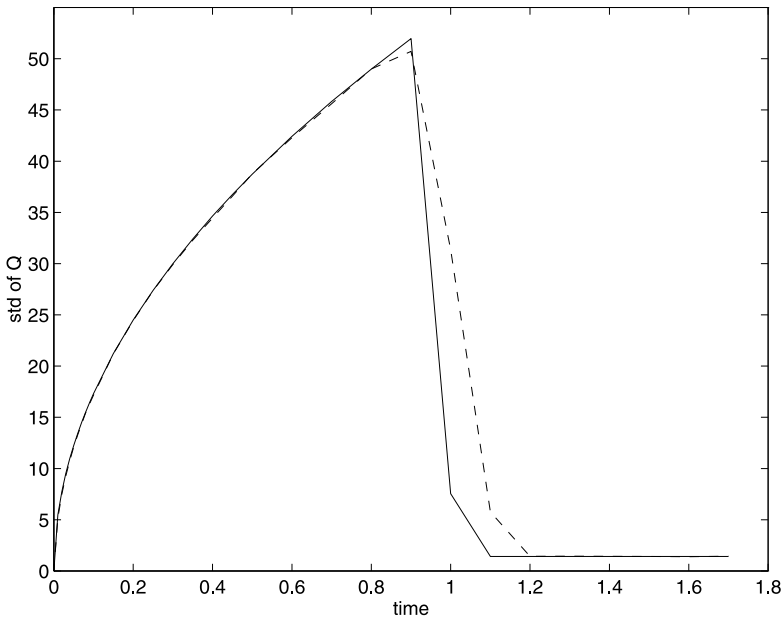
To be specific, let $Q(0) = 0$. Fluid and diffusion limits for the system defined by (11.9) are given by (4.7), (7.4) and (8.3), with

$$\lambda(\xi) = 2 \cdot \left\{ 1 - \frac{[\xi - (1 - \varepsilon)]^+}{\varepsilon} \right\}^+, \qquad \mu = 1,\ q(0) = 0.$$

Analysis of Figure 11 reveals that our approximations provide good estimators for $\mathbf{E}Q(\cdot)$ and $\operatorname{Var} Q(\cdot)$ during both the transient and the stationary phases.



(a)



(b)

FIG. 11. *Finite buffer model. Fluid and diffusion approximations versus simulation results. The solid lines are computed from fluid and diffusion approximations for the queueing system given by* (11.9), *with* $\varepsilon = 0.004$ *and* $n = 1000$. *The dashed lines are computed from* 10,000 *simulations of the original finite buffer system with* $\lambda = 2000$, $\mu = 1000$ *and* $C = 1000$: (a) $\mathbf{E}Q$; (b) $\boldsymbol{\sigma}Q$.

The histograms in Figure 12 demonstrate that the distribution of Q is well approximated by the normal distribution arising from our diffusion limit over a wide range of t. In particular, chi-square and Kolmogorov–Smirnov tests show that for $t = 0.5$ and $t = 0.8$, the empirical distributions fit the approximating normal distributions with a significance level of 0.95. However, for larger t (when the system operates within the stationary phase), we cannot expect a good fit between the empirical distribution of Q and the approximating normal distribution provided by our diffusion approximations. Indeed, even for $t = 0.9$ [see Figure 12d], the tail of the empirical density does not fit the approximating normal density. To understand this phenomenon, note that the discrete asymmetric stationary distribution of $M/M/(C + 1)$ is given by

$$p_k = \frac{1-\rho}{1-\rho^{C+2}}\rho^k, \qquad \rho = \frac{\lambda}{\mu}, \qquad k = 0, \ldots, C+1; \tag{11.10}$$

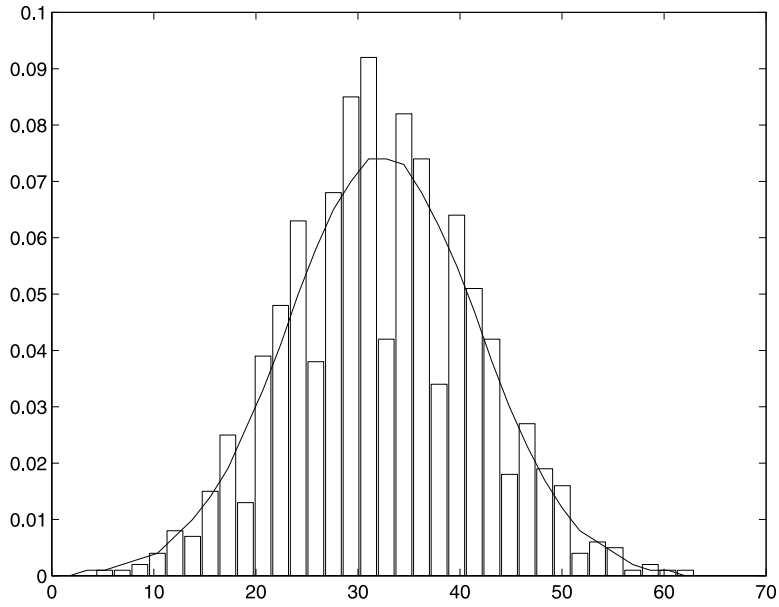
it is asymmetric, hence it cannot be approximated by a symmetric normal distribution. A rough explanation for this bad fit is as follows. When the system operates within the stationary phase, the free space in the buffer is no longer of order n. (For our case, $\rho = 2$ and, hence, the buffer is almost full within the stationary phase; see Figure 11.) To investigate queues with smaller buffers, we must use models and, respectively, diffusion approximations with additional reflection boundary at $\xi = C$. This reflection boundary introduces asymmetric distributions that would fit (11.10). (See, e.g., [32] and [18].)

Figure 13 exhibits a comparison between the finite buffer queue given by (11.8) and the approximations for the state-dependent queue given by (11.9), both with $n = 10$. We can see that for relatively small buffers ($\lambda = 20$, $\mu = 10$ and $C = 10$), our approximations are less satisfactory.

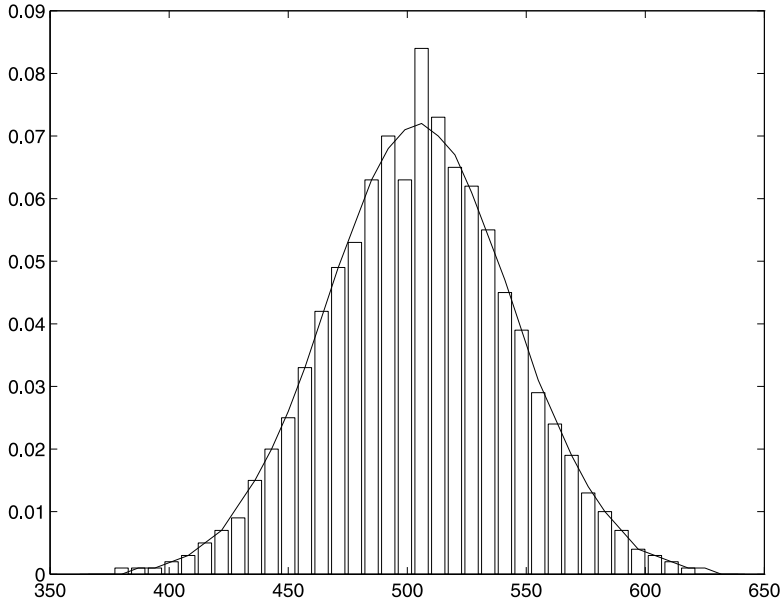
11.3. *Numerical comparison of different rescaling procedures.* The issue of different rescaling procedures was addressed in detail by Mandelbaum and Pats [59]. A summary of [59] is required to motivate the numerical examples presented below. Assume, instead of (7.3), that for some $\alpha \geq 0$,

$$\begin{aligned}
&\sqrt{n}\left(\frac{\lambda^n(n^\alpha\xi)}{n} - \lambda(\xi)\right) \to f_\lambda(\xi), \\
&\sqrt{n}\left(\frac{\mu^n(n^\alpha\xi)}{n} - \mu(\xi)\right) \to f_\mu(\xi), \\
&\sqrt{n}\,(P^n(n^\alpha\xi) - P(\xi)) \to f_P(\xi), \quad \text{u.o.c.},
\end{aligned} \tag{11.11}$$

when $n \uparrow \infty$. Our limit theorems correspond to $\alpha = 1$. Alternative rescaling procedures were considered by Yamada: the case $\alpha = 0$ was treated in [82], where the diffusion limit is of a Bessel type with a negative drift; the case $\alpha = 1/2$ was considered in [83], where the diffusion limit is a solution to a stochastic differential equation with *state*-dependent coefficients (while in our

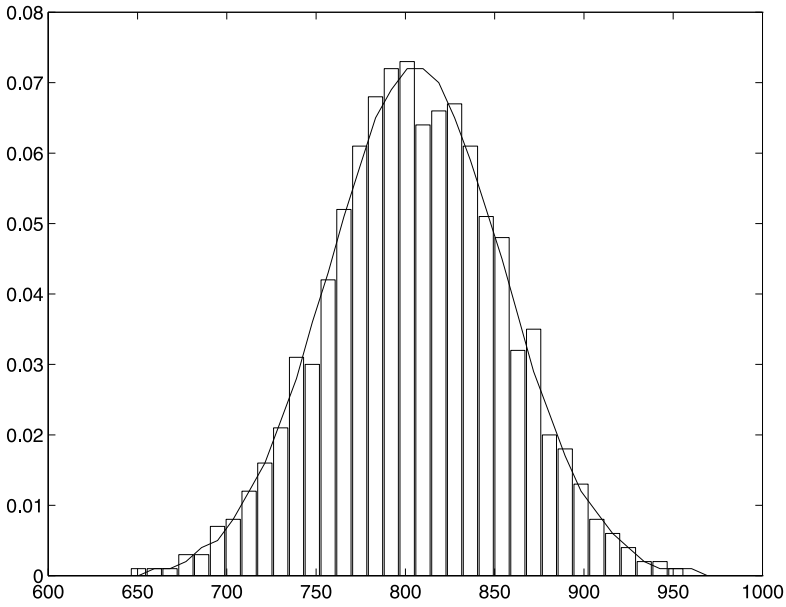


(a)

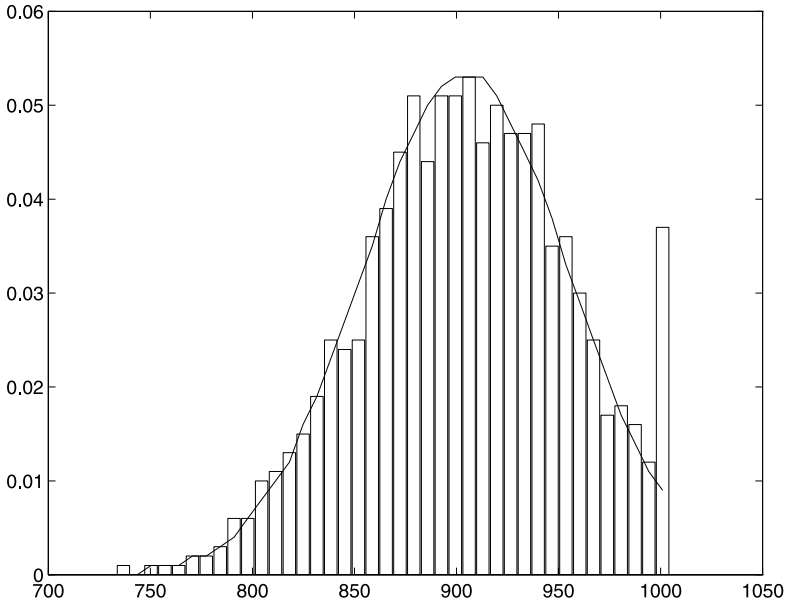


(b)

FIG. 12. *Finite buffer model. Comparison of the empirical distribution of* Q *at different times, computed from* 10,000 *simulations, with normal distributions provided by diffusion approximations*: (a) $t = 0.03, \mathcal{N}(31.6, 9.23)$; (b) $t = 0.5, \mathcal{N}(502, 38.8)$; (c) $t = 0.8, \mathcal{N}(802, 49)$; (d) $t = 0.9, \mathcal{N}(901, 50.7)$.

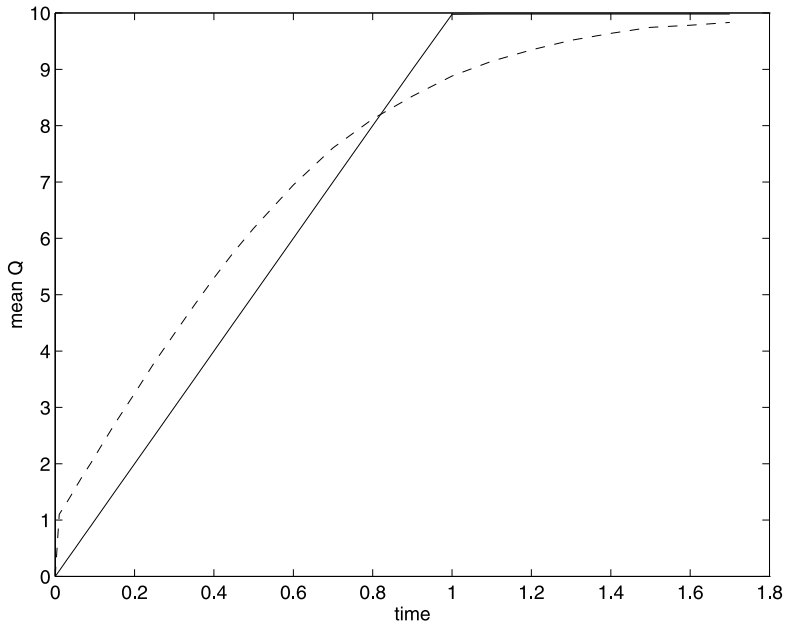


(c)

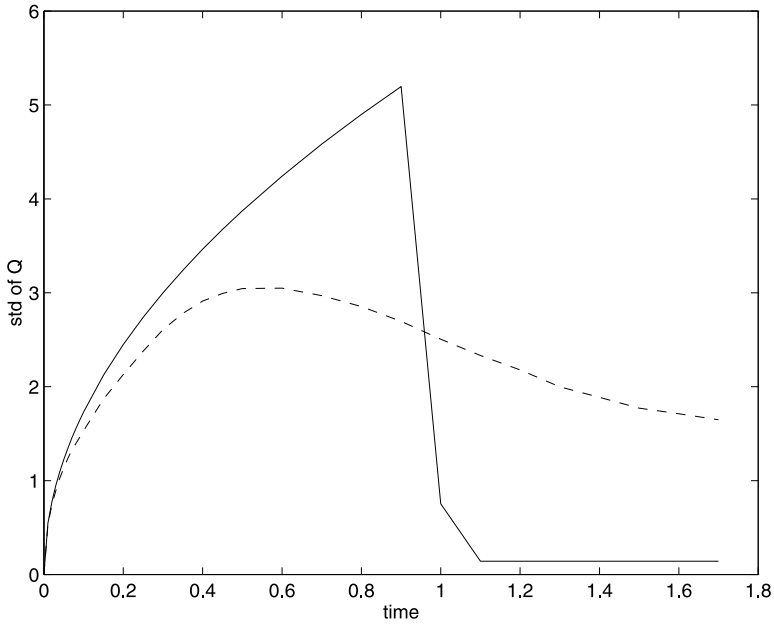


(d)

FIG. 12. *Continued*.



(a)



(b)

FIG. 13. *Finite buffer model. Comparison of diffusion approximations and simulation results. The solid lines are computed from fluid and diffusion approximations for the queueing system with* $\mu = 1$ *and* λ *given by* (11.9), *with* $\varepsilon = 0.004$, $n = 10$. *The dashed lines are computed from* 10,000 *simulations of the finite buffer system, with* $\lambda = 20$, $\mu = 10$ *and* $C = 10$: (a) $\mathbf{E}Q$; (b) $\boldsymbol{\sigma}Q$.

case, coefficients are *time*-dependent). The fluid limits vanish in both [82] and [83].

To recapitulate, our approach leads to second-order approximations for queueing processes: fluid limits provide approximations for the actual values of queues, while diffusion limits provide approximations for their fluctuations from the fluid limit. If fluid limits happen to vanish (operating within the critically loaded region), then

$$Q^n/\sqrt{n} \xrightarrow{d} V \tag{11.12}$$

and the three rescaling approaches above provide approximations for systems in which arrival and service rates are sensitive to fluctuations of queues: to small $[\alpha = 1,\ \mathscr{O}(n^{-1/2}V)]$, medium $[\alpha = 1/2,\ \mathscr{O}(V)]$ or large fluctuations $[\alpha = 0,\ \mathscr{O}(n^{1/2}V)]$.

We compare the three types of rescaling, $\alpha = 0, 1/2, 1$, by applying them to a single queueing system. Consider a sequence $M_\xi^n/M_\xi^n/1,\ n = 1, 2, \ldots,$ with arrival and service rates given by

$$\lambda^n(Q^n) = b^n + c^n \cdot (Q^n \wedge \delta^n), \qquad \mu^n(Q^n) = \beta^n + \gamma^n \cdot (Q^n \wedge \delta^n), \tag{11.13}$$

where $b^n, c^n, \delta^n, \beta^n, \gamma^n$ are positive constants, $c^n \le \gamma^n$. The following possible interpretations for the nth system were proposed in [59]:

1. Service is provided simultaneously by δ^n servers (each at a rate γ^n) and by a processor-shared server (at a rate β^n). The arrival process consists of exogenous arrivals (rate b^n) and served customers that leave for a while, then return for rework with probability $c^n/\gamma^n \le 1$. (The time until their return is assumed short enough that the queue does not change much, and long enough that they are independent of exogenous arrivals.) This is a possible model for some human-service systems.
2. Service is provided by a single server at a rate that increases with queue length, but only up to an exhaustion level $\beta^n + \gamma^n \cdot \delta^n$. Service rates, which increase with queue length, arise naturally in systems with reneging. (These are queues in which a customer is lost when its sojourn time reaches an individual random deadline [17]. In line with this interpretation, γ^n is the reneging rate.) Arrival rates which increase with queue length describe a possible scenario where a long queue attracts customers by being a source of information on service value.

Assume that $Q_0^n = 0$. The following three examples exhibit different diffusion limits V for different choices of parameters in (11.13):

1. $\alpha = 1$: Let $b^n = \beta^n = nb$, $c^n = \gamma^n = c$ and $\delta^n = n\delta$. Then $V = \sqrt{2b}\, W + Y$.
2. $\alpha = 1/2$: Let $b^n = nb$, $\beta^n = nb + \sqrt{n}$, $c^n = \sqrt{n}\, c$, $\gamma^n = \sqrt{n}\, c + 1$ and $\delta^n = \sqrt{n}\, \delta$. Then

$$\begin{aligned} dV_t = &-\left[1 + (V_t \wedge \delta)\right] dt + \sqrt{b + c(V_t \wedge \delta)}\, dW_t^1 \\ &+ \sqrt{b + c(V_t \wedge \delta)}\, dW_t^2 + dY_t. \end{aligned}$$

3. $\alpha = 0$: Let $b^n = \beta^n = nb$, $c^n = nc$, $\gamma^n = nc + \sqrt{n}$ and $\delta^n = \delta$. Then

$$V_t = -\delta \cdot t + \sqrt{2c\delta + 2b}\, W_t + Y_t.$$

Here $b, c, \delta > 0$, W, W^1, W^2 are standard Brownian motions (W^1 and W^2 are independent) and Y is a normal reflection term. For all three examples, the fluid limit $q \equiv 0$ and $Q^n/\sqrt{n} \to_d V$.

In what follows, numerical estimations of the queueing and idle-time processes in the three systems are provided. To approximate the idle-time process $I^n(\cdot) = \int_0^{\cdot} I\{Q^n(s) = 0\}\, ds$, we used the relation $I^n \sim_d \sqrt{n}\, Y/\beta^n$. It is a consequence of (8.1), (11.3) and the fact that we deal with the critically loaded regime [see (10.2)].

The parameters chosen for these experiments are $b = c = 0.25$, $\delta = 1$, and $n = 100$.

CASE 1. $\alpha = 1$. Figure 14 exhibits data for mean values and variances of the queueing and the idle-time processes. The distributions, expectations and standard deviations of V and Y are taken from [30], [1] and [2]:

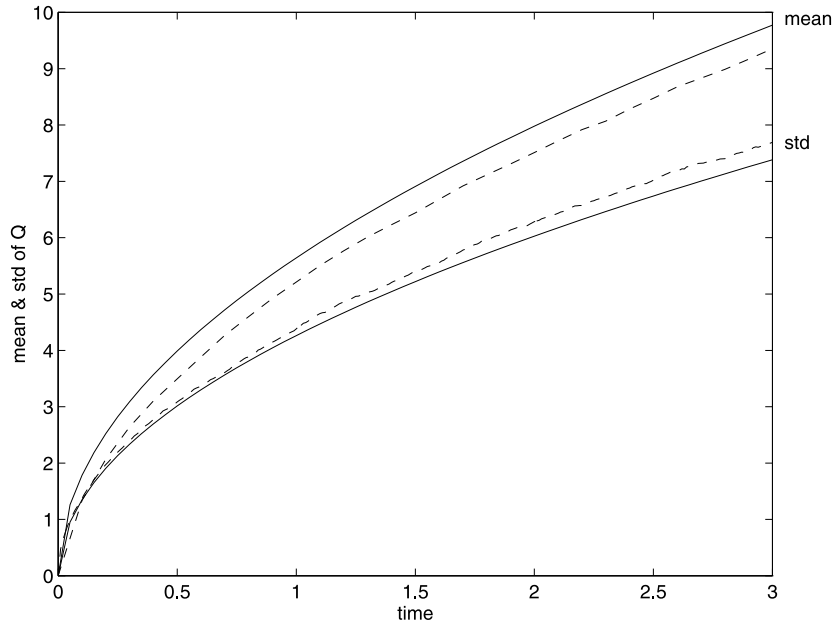
$$\begin{aligned} &\mathbf{P}\{V(t) \le z\} = \mathbf{P}\{Y(t) \le z\} = 1 - 2\Phi\left(\frac{-z}{\sqrt{2bt}}\right), \qquad z \ge 0, \\ &\mathbf{E}V(t) = \mathbf{E}Y(t) = 2\sqrt{bt/\pi}\,, \\ &\boldsymbol{\sigma}V(t) = \boldsymbol{\sigma}Y(t) = \sqrt{2bt(1 - 2/\pi)}\,, \qquad t \ge 0. \end{aligned} \tag{11.14}$$

In Figure 15, we compare the empirical distribution of $Q^n(2)/\sqrt{n}$, calculated from simulations, with the approximating distribution given by (11.14). As expected, the larger n gets, the better is the quality of the diffusion approximations. However, the results above demonstrate that our diffusion approximations also give reasonable estimations for relatively small queues when applied to critically loaded systems. Specifically, the relative error at $t = 3$ is 4.5% for $\mathbf{E}Q^n$, 4% for $\boldsymbol{\sigma}Q^n$, 3.6% for $\mathbf{E}I^n$ and 9% for $\boldsymbol{\sigma}I^n$.

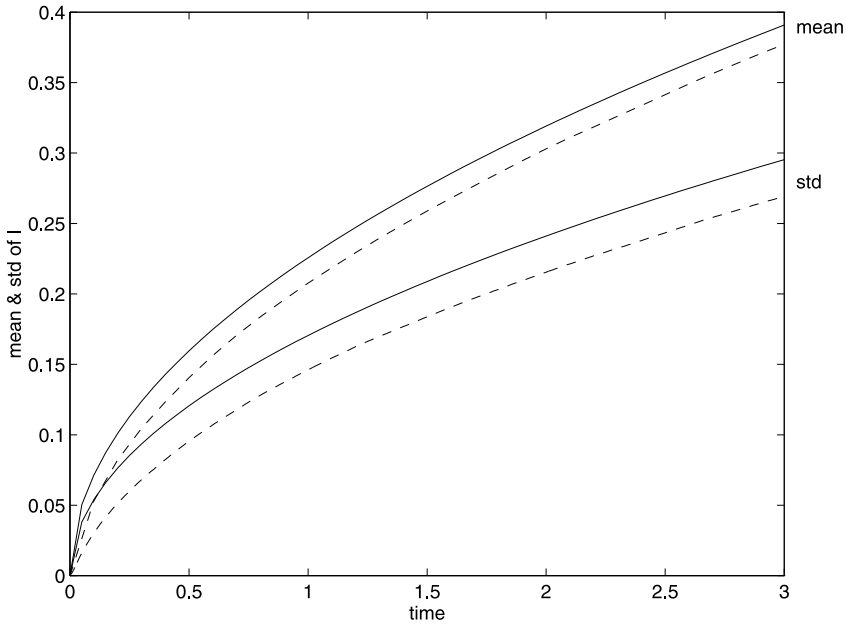
CASE 2. $\alpha = 1/2$. Figure 16 depicts mean values and variances of the queueing and idle-time processes. In Figure 17, we compare the empirical distribution of $Q^n(2)/\sqrt{n}$, calculated from simulations, with the approximating distribution obtained by numerical integration of the SDE for V. The relative error at $t = 2.5$ is 1.6% for $\mathbf{E}Q^n$, 15.8% for $\boldsymbol{\sigma}Q^n$, 0.8% for $\mathbf{E}I^n$ and 8.6% for $\boldsymbol{\sigma}I^n$.

CASE 3. $\alpha = 0$. Figure 18 exhibits data for mean values and variances of the queueing process Q^n for $n = 100$ and for $n = 10{,}000$. The diffusion approximations are computed from the equations (taken from [1] and [2])

$$\begin{aligned} \mathbf{E}V(t) &= 2^{-1} - (t+1)\left[1 - \Phi(\sqrt{t})\right] + \sqrt{t}\,\phi(\sqrt{t}), \\ \mathbf{E}V^2(t) &= 2^{-1} - (1 - 2t - t^2)\left[1 - \Phi(\sqrt{t})\right] + \sqrt{t}\,(1+t)\,\phi(\sqrt{t}), \end{aligned} \tag{11.15}$$

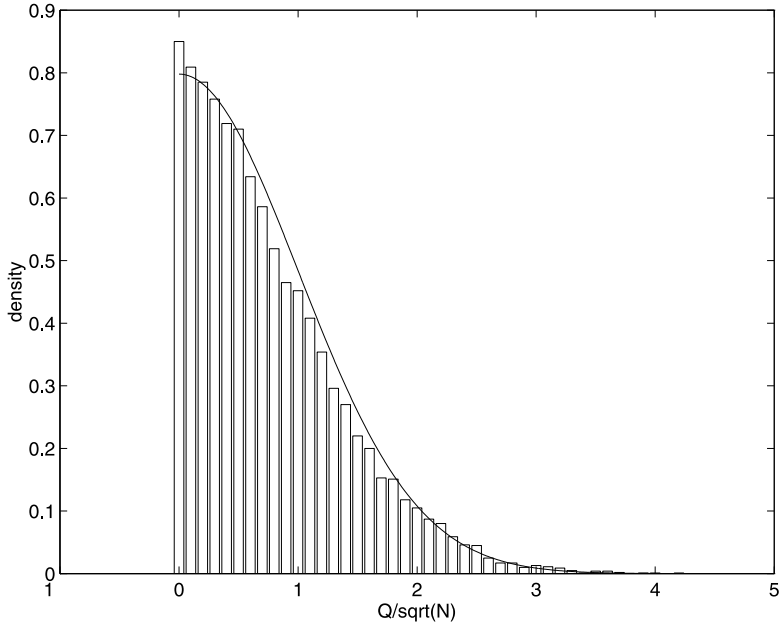


(a)

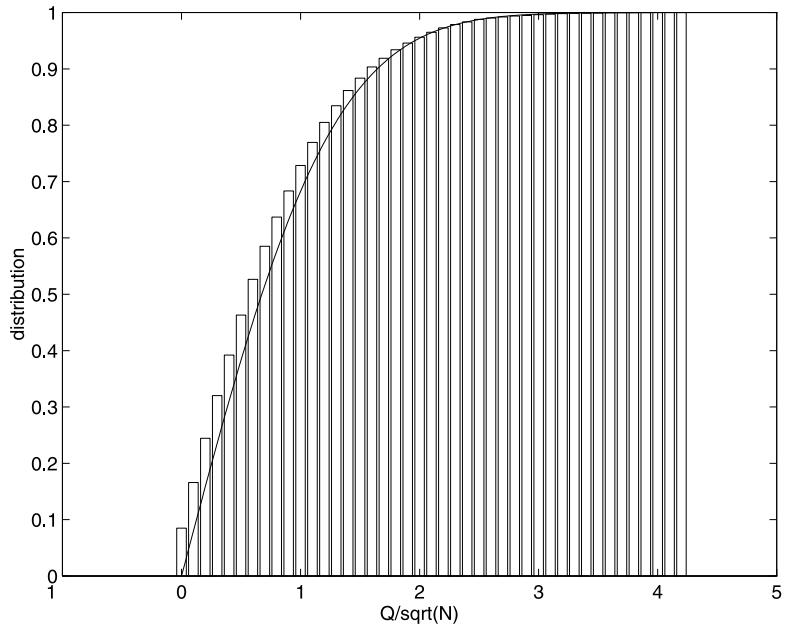


(b)

FIG. 14. $\alpha = 1$. *Diffusion approximations versus simulation results. The solid lines represent diffusion approximations. The dashed lines are computed from* 10,000 *simulations*: (a) $\mathbf{E}Q^n$, $\boldsymbol{\sigma}Q^n$; (b) $\mathbf{E}I^n$, $\boldsymbol{\sigma}I^n$.



(a)



(b)

FIG. 15. $\alpha = 1$. *Comparison of the empirical distribution of* $Q^n/\sqrt{n}$ *at* $t = 2$ *computed from* 10,000 *simulations with the theoretical distribution of RBM given by* (11.14): (a) *density of* $Q^n(2)/\sqrt{n}$; (b) *distribution of* $Q^n(2)/\sqrt{n}$.

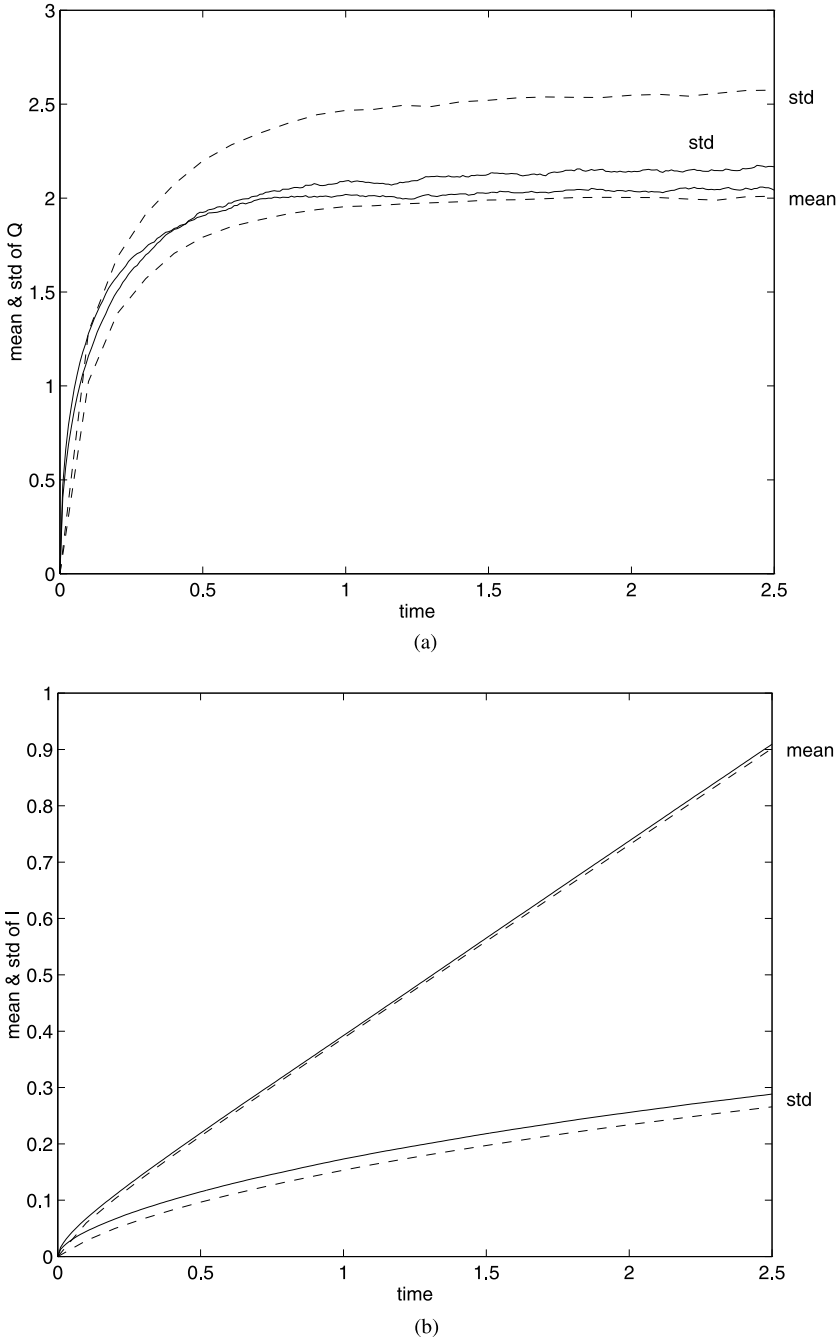
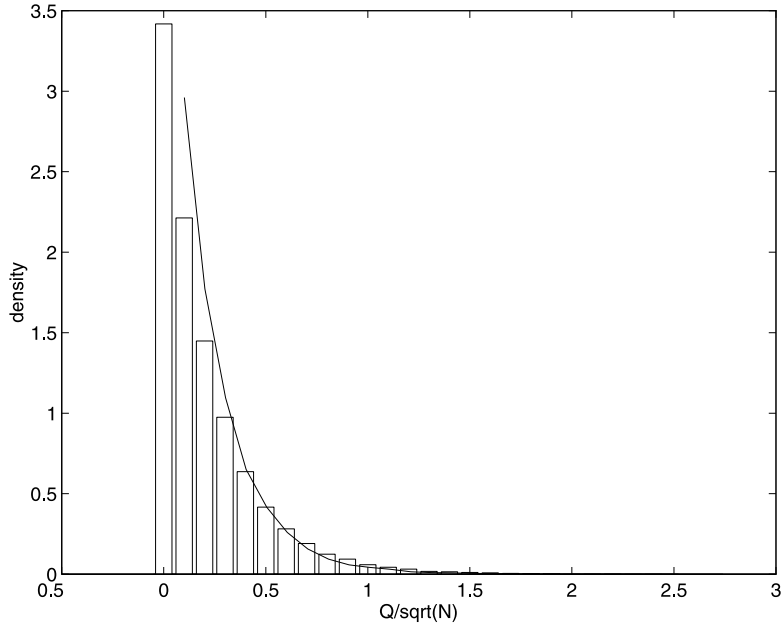
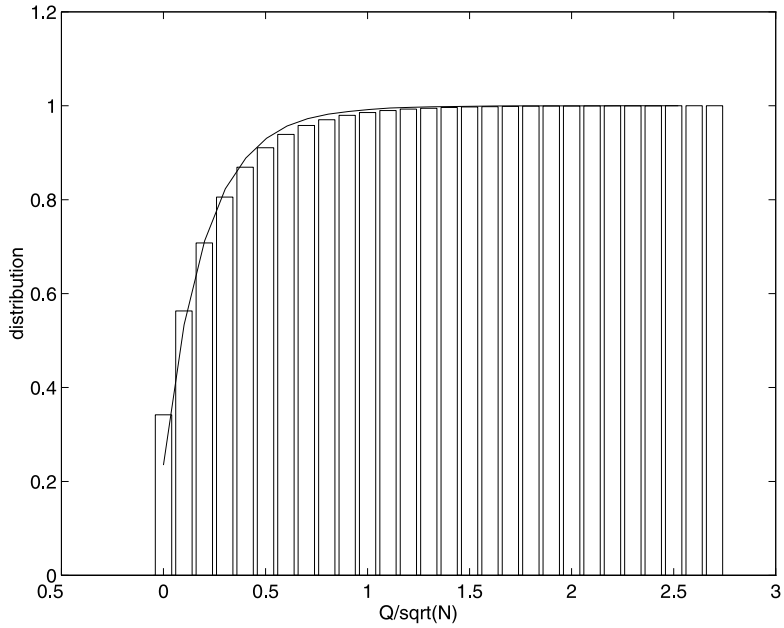


FIG. 16. $\alpha = 1/2$. *Diffusion approximations versus simulation results. The solid lines are computed from* 40,000 *numerical integrations of the corresponding SDE with step* $\Delta = 0.0001$. *The dashed lines are computed from* 100,000 *simulations*: (a) $\mathbf{E}Q^n$, $\boldsymbol{\sigma}Q^n$; (b) $\mathbf{E}I^n$, $\boldsymbol{\sigma}I^n$.

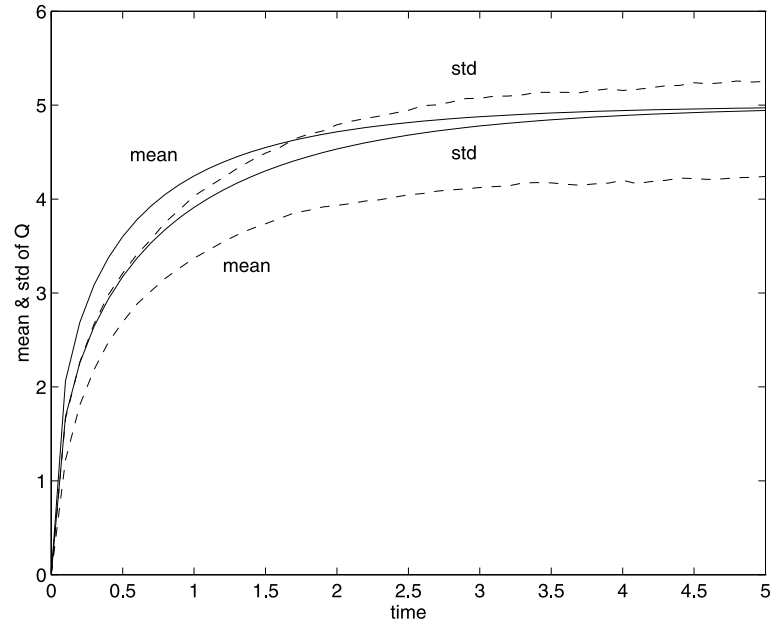


(a)

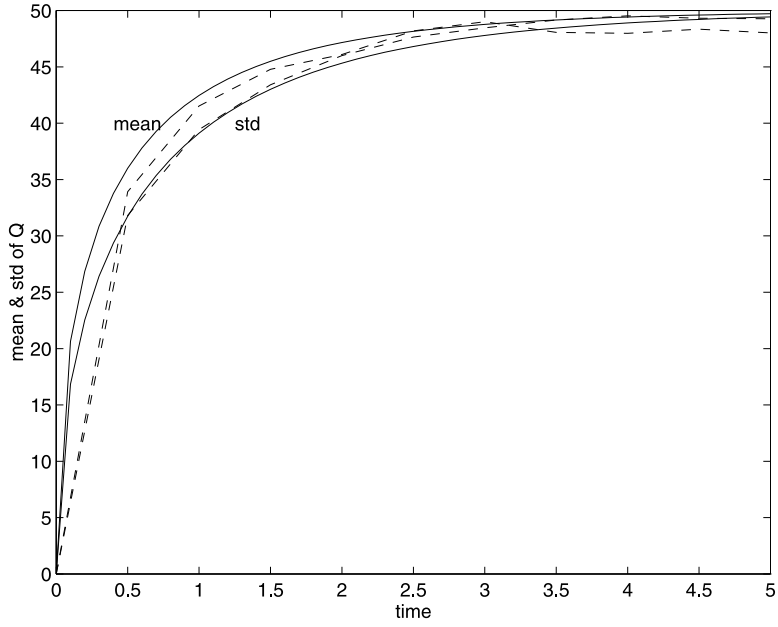


(b)

FIG. 17. $\alpha = 1/2$. *Comparison of the empirical distribution of* $Q^n/\sqrt{n}$ *at* $t = 2$ *computed from* 100,000 *simulations with the theoretical distribution of* V *computed from* 40,000 *integrations of the SDE with step* $\Delta = 0.0001$: (a) *density of* $Q^n(2)/\sqrt{n}$; (b) *distribution of* $Q^n(2)/\sqrt{n}$.



(a)



(b)

FIG. 18. $\alpha = 0$. *Comparison of diffusion approximations with simulation results for different* n. *The solid lines represent diffusion approximations* [*according to* (11.15)]. *The dashed lines are computed from simulations*: (a) $\mathbf{E}Q^n$, $\boldsymbol{\sigma}Q^n$; $n = 100$ (10,000 *simulations*); (b) $\mathbf{E}Q^n$, $\boldsymbol{\sigma}Q^n$; $n = 10{,}000$ (5000 *simulations*).

where ϕ and Φ are the standard normal density and distribution function, respectively. Comparison of Figure 18a and b demonstrates the quality improvement of fluid approximation as n increases. Figure 19 shows data for mean values and variances of the idle-time processes computed from simulations.

We conclude this example with some observations:

1. Analysis of Figures 14b, 16b and 19a reveals that our FCLT provides reasonable approximations for systems which operate under nonheavy traffic conditions. Specifically, the idle time permanently increases and comprises about 13, 36 and 31% of the total operation time in the first, second and third cases, respectively. Furthermore, queues in the systems considered above are relatively small.
2. Different rescaling procedures and the corresponding diffusion approximations can facilitate the design and analysis of queueing systems. For instance, recall the interpretation of μ^n in (11.13) as the service rate in a multiserver queue. Then, the examples above mainly differ by the number of servers relative to the queue size, which are $n:\sqrt{n}$, $\sqrt{n}:\sqrt{n}$ and $1:\sqrt{n}$ in examples 1, 2 and 3 respectively [the queue size is always of order $\sqrt{n}$ according to (11.12)]. Note that a comparison between these systems is meaningful in the sense that the potential arrival rates $b^n + c^n\delta^n$ and the total potential service rate $\beta^n + \gamma^n\delta^n$ are of order n in all three cases. Thus, the examples above present three different ways to allocate service capacity, n in total, among several servers. Analysis through Figures 14–19 gives rise to the following evaluation: (i) the magnitude of the queues is largest in Case 1 and smallest in Case 2; (ii) the coefficients of variations of queues are smallest in Case 1 and increase in the other two cases; (iii) the stationary distributions for queues exists in Cases 2 and 3, but not in Case 1; (iv) idle times reach the largest values in Case 1 and the smallest in Case 3.

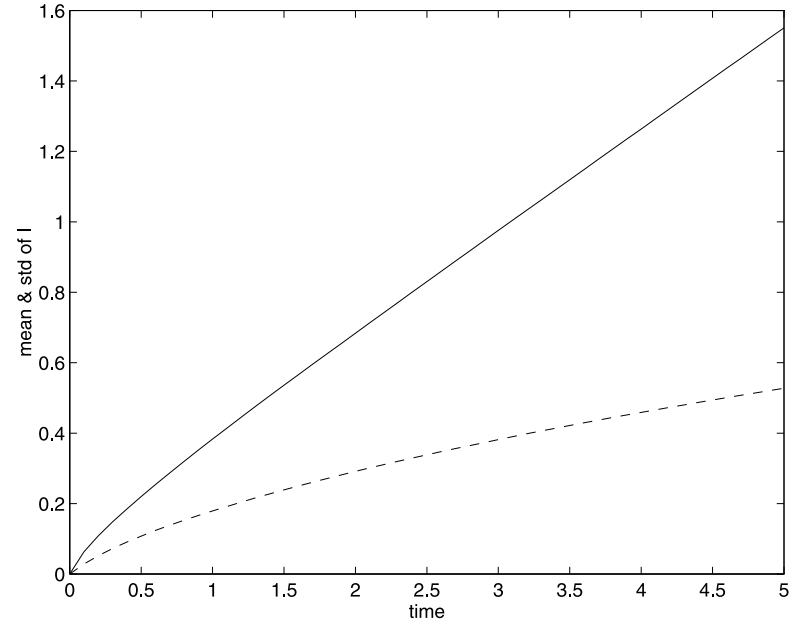
12. Proof of the martingale representation. This section is devoted to the proof of Lemma 3.9.

Our arguments, based on a multiparameter time change, are a straightforward adaptation of those given in Ethier and Kurtz [25], Chapter 6, Section 2, Kurtz [51] and Massey and Whitt [60], Lemma 2.2. Therefore, we omit some technical details.

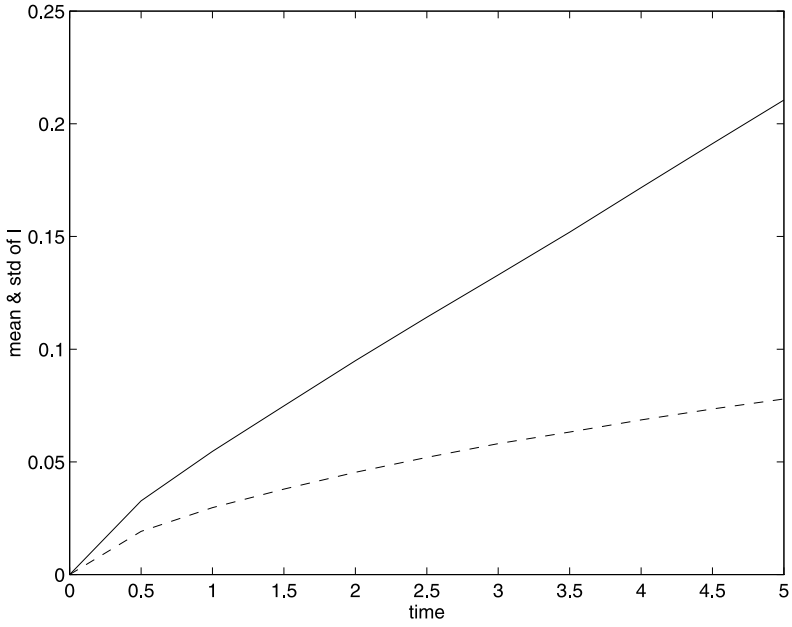
To simplify the presentation, we carry out the proof under stronger conditions than Main Assumption (M2) in Section 2. Namely, we assume that λ and μ are bounded: there exists a constant L_1 such that

$$|\lambda(\xi)|_\infty \vee |\mu(\xi)|_\infty < L_1, \qquad \xi \in \mathbb{R}_+^K.$$

To cover the general case, we can easily modify our arguments, taking into account Proposition 13.4 [specifically, inequality (13.5)] and Main Assumption (M2).



(a)



(b)

FIG. 19. $\alpha = 0$. *Simulation. The solid line is* $\mathbf{E}I^n$ *and the dashed line is* $\boldsymbol{\sigma}I^n$: (a) $n = 100$, (b) $n = 10{,}000$.

PROVING THE MARTINGALE PROPERTY OF M^a AND M^d. Introduce the following collection of σ-fields, for $k = 1, \ldots, K$, $t \geq 0$:

$$\mathscr{H}_k^+(t) = \sigma\big(N_k^+(u),\ 0 \leq u \leq t\big), \qquad \mathscr{H}_k^-(t) = \sigma\big(N_k^-(u),\ 0 \leq u \leq t\big),$$

$$\mathscr{G}_k(t) = \sigma\big(U_k\big[l \wedge N_k^-(t)\big],\ l = 0, 1, \ldots\big); \qquad k = 1, \ldots, K,\ t \geq 0.$$

For each $s = (s_1, s_2, \ldots, s_{2K}) \in [0, \infty)^{2K}$, let $\mathscr{H}(s)$ denote the σ-field given by

$$\mathscr{H}(s) = \bigvee_{k=1}^{K} \big[\mathscr{H}_k^+(s_{2k-1}) \vee \mathscr{H}_k^-(s_{2k}) \vee \mathscr{G}_k(s_{2k})\big] \vee \sigma(\mathscr{N}),$$

where $\mathscr{N}$ is the collection of all null sets in $\Im$. Finally, without loss of generality, assume that $\mathscr{H} = \{\mathscr{H}(s),\ s \in \mathbb{R}_+^K\}$ is right continuous.

Repeating the arguments of Theorem 2.2 in [25], Section 2, Chapter 6, which are based on the uniqueness of the solution to (2.1)–(2.5), leads to the following assertion: for all $t \geq 0$, the random vector $\tau = (\tau_1^+(t), \tau_1^-(t), \ldots, \tau_K^+(t), \tau_K^-(t))$, with

$$\tau_k^+(t) = \int_0^t \lambda_k(Q(s))\, ds, \qquad \tau_k^-(t) = \int_0^t \mu_k(Q(s))\, ds,$$

is a multiparameter $\mathscr{H}$-stopping point ([25], Section 8, Chapter 2).

Put $\mathscr{F}(t) = \mathscr{H}(\tau(t))$ and $\mathbf{F} = \{\mathscr{F}(t),\ t \geq 0\}$. In view of Main Assumption (M1) in Section 2, we infer that

$$\big(N_1^+(s_1) - s_1, N_1^-(s_2) - s_2, \ldots, N_K^+(s_{2K-1}) - s_{2k-1}, N_K^-(s_{2K}) - s_{2K}\big)$$

is a multiparameter martingale with respect to $\mathscr{H}$. Then the optional sampling theorem ([25], Theorem 8.7, Chapter 2) implies that M^a is a vector-valued $\mathbf{F}$-martingale, being a multiparameter time change of a multiparameter martingale. Moreover, M^a is also locally square integrable, because λ is bounded [69].

We proceed with the proof of the martingale property for M^d. For the same reasons as presented above, the process M^s, given by

$$M^s = S - \hat{S}, \qquad \hat{S}(t) = \int_0^t \mu(Q(u))\, du, \qquad t \geq 0,$$

is a vector-valued $\mathbf{F}$-martingale. Returning to (2.1)–(2.5), we have $Q(t) \in \mathscr{F}(t)$. Therefore, the integrand in (2.4) is a predictable process and thus M^d is a (locally square integrable) $\mathbf{F}$-martingale (see the integration theorem in [11], Theorem T8, page 27).

PROVING THE MARTINGALE PROPERTY OF M_f. Introduce the processes $M_{jk}^l = \{M_{jk}^l(t),\ t \geq 0\}$, $l = 1, 2, 3$, by

$$M_{jk}^1(t) = \int_0^t p_{jk}(Q(u-))\, dS_j(u) - \int_0^t p_{jk}(Q(u-))\, \mu(Q(u))\, du,$$

$$M_{jk}^2(t) = \int_0^t 1\big\{U_j\big[S_j(u)\big] \in \pi_{jk}(Q(u-))\big\}\, dS_j(u) - \int_0^t p_{jk}(Q(u-))\, \mu(Q(u))\, du,$$

$$M_{jk}^3(t) = \int_0^t 1\{U_j[S_j(u)] \in \pi_{jk}(Q(u-))\}\, dS_j(u) - \int_0^t p_{jk}(Q(u-))\, dS_j(u), \qquad j,k = 1,\dots,K.$$

Again, the integration theorem [11] implies, first, that M_{jk}^1 is a martingale (for all j,k) and, second, that to prove the martingale property for M^f, it is sufficient to show that M_{jk}^2, $j,k = 1,\dots,K$, are martingales. Since $M_{jk}^2 = M_{jk}^1 + M_{jk}^3$, we can see that the following lemma completes the proof:

LEMMA 12.1. *The processes* M_{jk}^3, $j,k = 1,\dots,K$, *are* $\mathbf{F}$*-martingales.*

PROOF. Our proof is similar to that of Lemma 2.2 in [60]. Fix any j,k and t,t_0 such that $t > t_0 \geq 0$. Denote by $\tilde{S}_j[l]$ the moment of the l-jump of S_j, $l = 1,2,\dots$. From (2.1)–(2.5) it follows that $M_{jk}^3(t) \in \mathscr{F}(t)$. Then

$$\begin{aligned}\mathbf{E}[M_{jk}^3(t)|\mathscr{F}(t_0)] &= M_{jk}^3(t_0) + \mathbf{E}\Big[\int_{(t_0,t]} [1\{U_j[S_j(u)] \in \pi_{jk}(Q(u-))\} \\ &\qquad - p_{jk}(Q(u-))]\, dS_j(u)|\mathscr{F}(t_0)\Big] \\ &= M_{jk}^3(t_0) + \sum_{l=1}^{\infty} \mathbf{E}\big[\big[1\{U_j[S_j(t_0)+l] \in \pi_{jk}(Q(s_{lj}-))\} \\ &\qquad - p_{jk}(Q(s_{lj}-))\big]1\{s_{lj} \leq t\}|\mathscr{F}(t_0)\big] \\ &= M_{jk}^3(t_0),\end{aligned}$$

where $s_{lj} \triangleq \tilde{S}_j[S_j(t_0)+l]$. The last equality is a consequence of the following statements, which are easily verified:

(i) The random variable $U_j[S_j(t_0)+l]$ is independent of $\mathscr{F}(t_0)$ for all $l \geq 1$ and $t_0 \geq 0$.

(ii) The random variables $U_j[S_j(t_0)+l]$ and s_{lj} are independent for all $l \geq 1$ and $t_0 \geq 0$.

(iii) The random variables $U_j[S_j(t_0)+l]$ and $Q(s_{lj}-)$ are independent for all $l \geq 1$ and $t_0 \geq 0$.

The proof of Lemma 3.9 is now complete. □

REMARK 12.2. The same arguments as above, which are based on a multiparameter time change and the optional sampling theorem, yield that

$$M_k^a M_l^d, \qquad k,l = 1,\dots,K,$$

are all purely discontinuous (locally square integrable) martingales. Hence, none of the processes M_k^a and M_l^d, $k,l = 1,\dots,K$, jump simultaneously (see [53], Theorem 1, page 49). A similar assertion is valid for any pair M_k^a and M_l^a with $k \neq l$, $k,l = 1,\dots,K$.

13. Proof of FLLN. To simplify the presentation we first carry out this proof under Assumptions B, presented momentarily, which are stronger than Assumptions A in Section 4. Commentary on the general case is provided in Section 13.1.

ASSUMPTIONS B.

(B1) Assume that

$$\frac{1}{n}\lambda^n(n\xi) = \lambda(\xi), \qquad \frac{1}{n}\mu^n(n\xi) = \mu(\xi), \qquad P^n(n\xi) = P(\xi), \qquad \text{u.o.c.},$$

for all $n = 1, 2, \ldots$, where λ, μ and P are given vector- and matrix-valued globally Lipschitz bounded functions and $\sup_{\xi \in \mathbb{R}_+^K} r(P(\xi)) < 1$.

(B2) Assume that $q^n(0) \to_p q(0)$, where $q(0) \in \mathbb{R}_+^K$ is a given deterministic vector.

13.1. *Proofs under Assumptions B. Existence and uniqueness* for the solution to (4.7) follow from results in Dupuis and Ishii ([23], Section 5; specifically, see Theorem 5.1 and Corollary 5.2). It turns out that in the case when P is constant, the proof of uniqueness for DEs with reflection amounts to combining Gronwall's inequality with the Lipschitz property of the oblique reflection operator [13, 46]. However, Example 4.1 and Proposition 4.1 from [22] show that, when P is state-dependent, the corresponding reflection operator need not be Lipschitz continuous. (For the notion of reflection operators, see Remark B.2.) This suggests that many of the standard tools cannot be used to establish existence and uniqueness for (4.7). Appropriate methods for treating existence and uniqueness have been developed in [21], motivated by nonlinear partial DEs.

Convergence of q^n is based on the following

LEMMA 13.1. *The sequence* $\{\alpha^n\}$ *given by* (4.4) *satisfies* $\mathbf{P}\text{-}\lim_{n\uparrow\infty} \|\alpha^n\|_T = 0$.

PROOF. It is sufficient to show that

$$\mathbf{P} = \lim_{n\uparrow\infty} \|\alpha_k^n\|_T = 0, \qquad k = 1, \ldots, K.$$

By Lemma 3.9, $M_k^{a,n}$, $M_k^{d,n}$ and $M_k^{f,n}$ are locally square integrable (purely discontinuous) martingales. Since $\hat{A}_k^n$, $\hat{D}_k^n$ and $\hat{F}_k^n$ are continuous, we have ([53], Problem 3, page 60)

$$\langle M_k^{a,n} \rangle = \hat{A}_k^n, \qquad \langle M_k^{d,n} \rangle = \hat{D}_k^n, \qquad \langle M_k^{f,n} \rangle = \hat{F}_k^n.$$

Now Doob's inequality ([53, Section 9, Chapter 1]) implies that, for all $\varepsilon > 0$,

$$\mathbf{P}\left\{\frac{1}{n}\|M_k^{a,n}\|_T \geq \varepsilon\right\} = \mathbf{P}\{\|M_k^{a,n}\|_T \geq n\varepsilon\}$$

$$\leq \frac{1}{n^2\varepsilon^2}\mathbf{E}\langle M_k^{a,n}\rangle(T) = \frac{1}{n\varepsilon^2}\mathbf{E}\int_0^T \lambda(q^n(u))\, du,$$

and that similar inequalities hold for $\|M_k^{d,n}\|_T$ and $\|M_k^{f,n}\|_T$. The proof of Lemma 13.1 is thus complete, in view of the boundedness of λ and μ. □

Next, note that one can recast (4.2)–(4.5) and (4.7) and (4.8) in the form of a time-dependent reflection problem. [See Definition B.1 and (B.6).] Specifically, given existence and uniqueness of the solutions, we have

$$
\begin{aligned}
q^n &= \mathbf{\Phi}_{\mathscr{S}}^{\mathbf{R}}(x^n),\\
x^n(\cdot) &= q^n(0) + \frac{1}{n}\int_0^{\cdot} \theta^n(nq^n(u))\,du\\
&\quad - \frac{1}{n}\int_0^{\cdot} \left[\left[P^n(nq^n(u-))\right]^T - P^T(q(u))\right]\\
&\qquad \times I\{q^n(u) = 0\}\mu^n(nq^n(u))\,du + \alpha^n(\cdot)
\end{aligned}
\tag{13.2}
$$

and

$$
q = \mathbf{\Phi}_{\mathscr{S}}^{\mathbf{R}}(x), \qquad x(\cdot) = q(0) + \int_0^{\cdot} \theta(q(u))\,du,
\tag{13.3}
$$

where $\mathscr{S} \triangleq \mathbb{R}_+^K$ and $R(\cdot) = [I - P^T(q(\cdot))]$.

Now, subtracting the equation for q in (13.3) from the equation for q^n in (13.2) and using the Lipschitz properties of the time-dependent reflection (see Theorem B.1), θ and P, we obtain

$$
\|q^n - q\|_t \le L_T\left(|q^n(0) - q(0)| + \|\alpha^n\|_t + \int_0^t \|q^n - q\|_u\,du\right), \qquad t \in [0,T],
$$

for all $T > 0$ and for some $L_T > 0$. Thus, the assertion of the theorem follows from Assumption (B2), Lemma 13.1 and Gronwall's inequality ([25], page 428).

13.2. *Proofs under Assumptions A.* To prove Theorem 4.6 under Assumptions A in Section 4, the proof above can be redone, taking into account the following proposition.

PROPOSITION 13.4. *Let Assumptions A be satisfied. Then there exist a sequence* $\{\overline{L}^n\}$ *of positive random variable and deterministic constants* $L_0 \ge 1$, $\gamma > 0$, *such that*

$$
\begin{aligned}
&|q^n(t)|_1 \le \overline{L}^n e^{\gamma t} - 1, \qquad t \ge 0,\ n = 1,2,\ldots,\\
&\mathbf{P}\{\overline{L}^n > l\} \le \frac{L_0}{l},\\
&\mathbf{E}|q^n(t)|_1 \le L_0 e^{\gamma t} - 1, \qquad t \ge 0,\ n = 1,2,\ldots .
\end{aligned}
\tag{13.5}
$$

PROOF. The proof can be carried out by repeating the arguments in Kurtz ([50], Theorem 2.1). □

Roughly speaking, this lemma establishes probabilistic bounds on q^n, and hence on $\lambda^n(q^n)/n$ and $\mu^n(q^n)/n$, $n = 1, 2, \ldots$, as well [due to Assumption (A2) in Section 4]. These bounds allow replication of our proof for bounded rates under only minor modifications.

14. Proof of FCLT. Here, we extend and generalize Krichagina [46] to cover networks with state-dependent routing. To simplify the representation, we write down the proof of a weaker version of Theorem 7.2. Namely, we establish the convergence of $\{V^n\}$ under the following conditions:

1. Assumptions B in Section 13 are satisfied.
2. λ, μ and P are continuously differentiable with globally Lipschitz derivatives.
3. $J^+(\cdot)$, $J^-(\cdot)$ and $J^0(\cdot)$ are constant during $[0, T]$.
4. $V^n(0) \to_d V(0)$, where $V(0)$ is a given random vector with $V_k(0) = 0$ for all $k \in J^-$.

In this case, f_λ, f_μ and f_P in (7.3) vanish. Then V is a solution to the following SDE with time-dependent reflection:

$$V = \mathbf{\Phi}_\Gamma^{\mathbf{R}}\left([I - I^-] \int_0^\cdot \tilde{R}^{-1}(t)\, dX(t) \right), \tag{14.1}$$

$$\begin{aligned} X(\cdot) = V(0) &+ \int_0^\cdot \partial\theta(q(t))V(t)\, dt - \int_0^\cdot \partial P^T(q(t)) \odot V(t)\, dy(t) \\ &+ \int_0^\cdot \Sigma^{1/2}(q(t))\, dW(t). \end{aligned} \tag{14.2}$$

Further, in view of Assumptions B, (7.12)–(7.18) take the form

$$V^n = \sqrt{n}\left[\mathbf{\Phi}_{\mathscr{S}}^{\mathbf{R}}\left(x + \frac{1}{\sqrt{n}} X^n \right) - \mathbf{\Phi}_{\mathscr{S}}^{\mathbf{R}}(x) \right], \tag{14.3}$$

where

$$X^n(\cdot) = V^n(0) + B_\theta^n(\cdot) - B_P^n(\cdot) + M^n(\cdot), \tag{14.4}$$

$$B_\theta^n(\cdot) = \sqrt{n} \int_0^\cdot \{\theta(q^n(u)) - \theta(q(u))\}\, du, \tag{14.5}$$

$$B_P^n(\cdot) = \sqrt{n} \int_0^\cdot \left[P^T(q^n(u)) - P^T(q(u)) \right] I\{q^n(u) = 0\} \mu(q^n(u))\, du, \tag{14.6}$$

$$M^n = \sqrt{n}\, \alpha^n. \tag{14.7}$$

To prove FCLT in the general case, one can use Proposition 13.4 and a standard cutoff (or localization) argument ([73], Section 11.1; see also [72] for technical aspects of this argument).

14.1. *The main steps.* We describe below the main steps of the proof. Proofs are provided in Section 14.2.

LEMMA 14.8. *There exists a unique strong solution V to* (14.1) *and* (14.2). *The process V is continuous and Markovian.*

LEMMA 14.9. *For sequence $\{M^n\}$ in* (14.7),

$$M^n \xrightarrow{d} \tilde{W}, \tag{14.10}$$

$$\tilde{W}(\cdot) = \int_0^{\cdot} \Sigma^{1/2}(q(t))\, dW(t). \tag{14.11}$$

Here q and Σ are given by (4.7) *and* (7.10), *respectively, and W is a standard $\mathbb{R}^K$-valued Brownian motion.*

LEMMA 14.12. *The sequence $\{V^n\}$ adheres to the compact containment condition*

$$\lim_{l \uparrow \infty} \limsup_n \mathbf{P}\{\|V^n\|_T > l\} = 0.$$

LEMMA 14.13. *The sequence $(V^n, X^n, B_\theta^n, B_P^n, M^n)$ is C-tight.*

LEMMA 14.14. *Let $(V, X, B_\theta, B_P, \tilde{W})$ be any weak limit of $(V^n, X^n, B_\theta^n, B_P^n, M^n)$. Then X satisfies* (14.2).

14.2. *Proofs.*

PROOF OF LEMMA 14.8. Rewrite (14.1) and (14.2) in the form ($t \in [0, T]$)

$$V = \mathbf{\Phi}_\Gamma^{\mathbf{R}}\big([\, I - I^- \,]\tilde{X}\big), \qquad \tilde{X}(0) = \tilde{R}^{-1}(0)V(0),$$

$$d\tilde{X}(t) = \tilde{R}^{-1}(t)\Big\{ g\Big[q(t), \mathbf{\Phi}_\Gamma^{\mathbf{R}}\big([\, I - I^- \,]\tilde{X}\big)(t)\Big]\, dt + \Sigma^{1/2}(q(t))\, dW(t)\Big\}, \tag{14.15}$$

where

$$g[\, \xi, \chi \,] = \partial\theta(\xi)\chi - \partial P^T(\xi) \odot \chi \cdot m(\xi), \qquad \xi \in \mathbb{R}_+^K,\ \chi \in \mathbb{R}^K.$$

Note that $\Sigma(q(\cdot))$ is a Lipschitz function. Now, since $\mathbf{\Phi}_\Gamma^{\mathbf{R}}$ is nonanticipating and Lipschitz (Theorem B.1), there exists a unique strong solution to (14.15) and hence to (14.1) and (14.2) as well. Moreover, the continuity and Markovian property of V immediately follow, in view of Propositions B.1 and B.2. Actually, weak existence follows from our proof of the FCLT, where it is shown that limit points of some C-tight sequence satisfy (14.1) and (14.2). □

PROOF OF LEMMA 14.9. Recall that $\{M^n\}$ is a sequence of locally square integrable martingales [see Lemma 3.9, (4.4) and (14.7)]. By Theorem 4 of Liptser and Shiryayev ([53], page 567), to establish (14.10) it is sufficient to show that

$$\langle M^n \rangle(t) \xrightarrow{p} \int_0^t \Sigma(q(u))\, du, \qquad t \in [0, T]. \tag{14.16}$$

We now proceed with the proof of (14.16). It follows from (7.10) that

$$\begin{aligned}&\int_0^{\cdot} \Sigma(q(u))\,du\\ &\quad = \operatorname{diag}\{c^1(\cdot)\} + \operatorname{diag}\{c^2(\cdot)\} + \operatorname{diag}\{c^{3,n}(\cdot)\}\\ &\qquad - \int_0^{\cdot} P^T(q(u))\,d\big(\operatorname{diag}\{c^2(u)\}\big) - \int_0^{\cdot} d\big(\operatorname{diag}\{c^2(u)\}\big)P(q(u)),\end{aligned} \tag{14.17}$$

$$\begin{aligned}&c^1(\cdot) = \int_0^{\cdot} \lambda(q(u))\,du, \qquad c^2(\cdot) = \int_0^{\cdot} \{\mu(q(u)) - m(q(u))\}\,du,\\ &c^3(\cdot) = \int_0^{\cdot} P^T(q(u))\{\mu(q(u)) - m(q(u))\}\,du.\end{aligned} \tag{14.18}$$

Next, in view of (3.3) and (14.7), we have

$$\begin{aligned}\langle M^n\rangle = \frac{1}{n}\big[&\langle M^{a,n}\rangle + \langle M^{a,n}, M^{f,n}\rangle - \langle M^{a,n}, M^{d,n}\rangle\\ &+\langle M^{f,n}, M^{a,n}\rangle + \langle M^{f,n}\rangle - \langle M^{f,n}, M^{d,n}\rangle\\ &-\langle M^{d,n}, M^{a,n}\rangle - \langle M^{d,n}, M^{f,n}\rangle + \langle M^{d,n}\rangle\big].\end{aligned}$$

Note that $M^{a,n}$, $M^{f,n}$ and $M^{d,n}$ are purely discontinuous martingales (see Theorem 3 in [53], page 41). Recall that for two purely discontinuous martingales M^1 and M^2, $\langle M^1, M^2\rangle_{\cdot}$ coincides with the compensator of $\sum_{u\le \cdot} \Delta M^1(u)\,\Delta M^2(u)$. (See, e.g., [53], Theorem 3, page 41, and [24], Theorem 9.40.) Then Remark 12.2 immediately implies that

$$\langle M^n\rangle = \frac{1}{n}\big[\langle M^{a,n}\rangle + \langle M^{f,n}\rangle + \langle M^{d,n}\rangle - \langle M^{f,n}, M^{d,n}\rangle - \langle M^{d,n}, M^{f,n}\rangle\big].$$

Continuing computations, one can easily obtain that

$$\begin{aligned}\langle M^n\rangle(\cdot) = {}&\operatorname{diag}\{c^{1,n}(\cdot)\} + \operatorname{diag}\{c^{2,n}(\cdot)\} + \operatorname{diag}\{c^{3,n}(\cdot)\}\\ &- \int_0^{\cdot} P^T(q^n(u))\,d\big(\operatorname{diag}\{c^{2,n}(u)\}\big)\\ &- \int_0^{\cdot} d\big(\operatorname{diag}\{c^{2,n}(u)\}\big)P(q^n(u)),\end{aligned} \tag{14.19}$$

$$\begin{aligned}&c^{1,n}(\cdot) = \int_0^{\cdot} \lambda(q^n(u))\,du,\\ &c^{2,n}(\cdot) = \int_0^{\cdot} I\{q^n(u-) > 0\}\mu(q^n(u))\,du,\\ &c^{3,n}(\cdot) = \int_0^{\cdot} P^T(q^n(u))\,I\{q^n(u-) > 0\}\mu(q^n(u))\,du.\end{aligned} \tag{14.20}$$

The proof of (14.10) is complete once we show that each of the five terms on the right-hand side of (14.19) converges uniformly over $[0, T]$ in probability to

the corresponding term on the right-hand side of (14.17). To establish this convergence, it is sufficient to show that, for $i = 1, 2, 3$,

$$c^{i,n}(\cdot) \xrightarrow{p} c^i(\cdot), \qquad n \uparrow \infty. \tag{14.21}$$

First, since λ is Lipschitz, $\{c^{1,n}\}$ converges by FLLN. Now, (13.2) and Assumptions B imply

$$q^n = \mathbf{\Phi}^{\mathbf{R}}_{\mathcal{G}}(x^n), \qquad x^n(\cdot) = q^n(0) + \int_0^{\cdot} \theta(q^n(u))\, du - \epsilon^n(\cdot) + \alpha^n(\cdot),$$

where

$$\begin{aligned} \epsilon^n(\cdot) &= \int_0^{\cdot} \left[P^T(q^n(u-)) - P^T(q(u))\right] dy^n, \\ y^n(\cdot) &= \int_0^{\cdot} I\{q^n(u) = 0\}\mu(q^n(u))\, du. \end{aligned} \tag{14.22}$$

By properties of the Stiltjes integral, FLLN implies that $\epsilon^n \to_p 0$ u.o.c. Then, by Lemma 13.1 and by the Lipschitz property of time-dependent reflection, we have

$$\int_0^t \left[I - P^T(q(u))\right] dy^n(u) \xrightarrow{p} \int_0^t \left[I - P^T(q(u))\right] dy(u),$$

where y is given by (4.7). In view of (C.2), simple arguments from calculus imply that

$$y^n \xrightarrow{p} y. \tag{14.23}$$

Then the convergence in (14.21) for $i = 2, 3$ follows. □

PROOF OF LEMMA 14.12. During this proof we use positive constants C_i, $i = 1, 2, 3$. Explicit expressions for these constants are of no significance and therefore are not given.

By the Lipschitz property of $\mathbf{\Phi}^{\mathbf{R}}_{\mathcal{G}}$, it follows from (14.3) that

$$\|V^n\|_t \le C_1 \|X^n\|_t, \qquad t \in [0, T].$$

Using Lipschitz properties of θ and P, we infer from (14.4)–(14.7) that

$$\|X^n\|_t \le |V^n(0)| + \|M^n\|_t + C_2 \int_0^t \|V^n\|_u\, du, \qquad t \in [0, T].$$

Combining the two inequalities above with Gronwall's lemma [25] implies:

$$\|V^n\|_T \le C_1(|V^n(0)| + \|M^n\|_T)\exp(C_3 T).$$

Now, $\{V^n(0)\}$ and $\{M^n\}$ converge weakly by conditions of this theorem and by Lemma 14.9, respectively. The proof of this lemma is thus complete. □

PROOF OF LEMMA 14.13. Recall (7.19), (7.20) and considerations thereafter. It was explained there that the assertion of Lemma 14.13 would follow if C-tightness of $\{X^n\}$ is established. In view of Lemma 14.9, it suffices to show

that $\{B_\theta^n\}$ and $\{B_P^n\}$, given by (14.5) and (14.6), respectively, are tight in $C^K[0,T]$. We restrict our attention to $\{B_\theta^n\}$, since a proof of tightness for $\{B_P^n\}$ is completely analogous. Since θ is Lipschitz, (14.5) yields

$$|B_\theta^n(t) - B_\theta^n(s)| \le C\|V^n\|_T(t-s), \qquad 0 \le t \le s \le T,$$

for some $C > 0$. Hence, tightness of $\{B_\theta^n\}$ follows from Lemma 14.12 (see, e.g., [38], Proposition VI.3.26). □

PROOF OF LEMMA 14.14. In view of Lemma 14.9, it suffices to show that

$$(14.24) \quad B_\theta^n \xrightarrow{d} \int_0^\cdot \partial\theta(q(t))V(t)\,dt, \qquad B_P^n \xrightarrow{d} \int_0^\cdot \partial P^T(q(t)) \odot V(t)\,dy(t).$$

To prove the first limiting relation in (14.24), note that

$$B_\theta^n(\cdot) = \int_0^\cdot f_\theta^n(t)V^n(t)\,dt, \qquad f_\theta^n(\cdot) = \int_0^1 \partial\theta\left(q(\cdot) + \frac{u}{\sqrt{n}}V^n(\cdot)\right)du.$$

Thus, combining FLLN, the bounded convergence theorem and the continuous mapping theorem implies the convergence for $\{B_\theta^n\}$. Analogously,

$$B_P^n(\cdot) = \int_0^\cdot f_P^n(t) \odot V^n(t)\,dy^n(t), \qquad f_P^n(\cdot) = \int_0^1 \partial P^T\left(q(\cdot) + \frac{u}{\sqrt{n}}V^n(\cdot)\right)du,$$

where y^n is given by (14.22). The proof is thus complete by (14.23) and by properties of the Stiltjes integral. □

15. Future research. Our efforts are currently directed toward covering discontinuous diffusion limits and approximating networks which are both state- and time-dependent. We outline below these two directions.

15.1. *M_1-convergence*. Discontinuous limits arise when the conditions of FCLT (Theorem 7.2) are relaxed. We describe below the general state of affairs. Formulations and proofs can be found in [65] and will appear in a complete form in a future paper [64]. As already pointed out, convergence to discontinuous limits holds in the M_1-topology. Within the context of a single station, this issue is considered in [59].

Recall that at each moment t, every station of the network belongs to one of the sets $J^+(t)$, $J^-(t)$ or $J^0(t)$; namely, it is overloaded, underloaded or critically loaded, respectively [see (6.2)]. Theorem 7.2 reveals that diffusion limits for stations in J^+ are general diffusion processes, in J^0 they are *reflected* diffusions and for stations in J^- they vanish. It is explained in Section 7 that a diffusion limit can jump with positive probability only at those times when the corresponding fluid limit switches from one region to another. (We say that there are *phase transitions* at those times.) To be more specific, we appeal to Mandelbaum and Massey [55], where a single time-dependent station is treated. The following observation is behind the similarity between the cases arising here and in [55]. Let q be the fluid limit for a

given state-dependent network: q is the unique solution to (4.7). Define

$$\tilde{\lambda}(t) \triangleq \lambda(q(t)) + P^T(q(t))\mu(q(t)),$$
$$\tilde{\mu}(t) \triangleq \mu(q(t)) + P^T(q(t))m(q(t)), \qquad t \geq 0,$$

where m is given by (6.1). Then q is the unique solution to the normal reflection problem

$$q(t) = q(0) + \int_0^t \big(\tilde{\lambda}(u) - \tilde{\mu}(u)\big)\, du + y(t) \geq 0, \qquad t \geq 0,$$
$$y \text{ is nondecreasing in each coordinate, } y(0) = 0,$$
$$\int_0^\infty 1^T[q(t) > 0]\, dy(t) = 0.$$

This equation has the form of a fluid approximation for a single time-dependent station. Now the results of [55] imply that the diffusion limit is discontinuous at t_0 (with positive probability) exactly in the following cases:

CASE 1. $t_0 > 0$. (i) $k \in J^+(t)$, $t \in [t_0 - \varepsilon, t_0)$, $k \notin J^+(t_0)$; (ii) there exist sequences $t^n \uparrow t_0$ and $\tilde{t}^n \downarrow t_0$ such that, for some k and $\varepsilon > 0$,

$$k \in J^+(t) \cup J^0(t), \qquad t \in [t_0 - \varepsilon, t_0) \quad \text{and}$$
$$k \in J^0(t^n), \qquad k \in J^-(\tilde{t}^n), \qquad n \uparrow \infty.$$

CASE 2. $t_0 = 0$. $k \in J^-(t)$, $t \in [0, \varepsilon)$, $V_k(0) \not\equiv 0$, for some k and $\varepsilon > 0$.

It turns out that even if only *some* coordinates of the fluid limit undergo phase transition, then discontinuities can arise at *all* coordinates of the diffusion limit.

The main idea of our FCLTs with discontinuous diffusion limits is to analyze V^n on different time scales (slowly varying time scale near points of phase transitions.) This random time change enables us to pick up the behavior of $\{V^n\}$ during short phases, which shrink under rescaling and give rise to the discontinuities. (For additional details, see [59], Section 4.5.)

15.2. *Time-dependent networks*. The results obtained in this paper provide insight into the nature of fluid and diffusion approximations for networks, which are state- and time-dependent simultaneously. For simplicity of presentation, we focus below on purely time-dependent networks. The general case can be treated similarly.

Consider a K-station network given by (2.1)–(2.5). Append to this network an additional station $(K + 1)$, which is disconnected from all other stations, and assume that $\lambda_{K+1} \equiv 1$ and $\mu_{K+1} \equiv 0$. Clearly, $Q_{K+1} = N^+_{K+1}$ (standard Poisson). Next, suppose that the arrival and service rates at stations $1, \ldots, K$, as well as the routing probabilities, depend only on Q_{K+1}.

Now, FLLN for Poisson processes implies that, under our rescaling [Assumption (A1) in Section 4], the fluid limit for $(K + 1)$th station is $q_{K+1}(t) =$

t, $t \geq 0$. This suggests that (4.7), which defines the fluid limit q for the original network, reduces to a *time-dependent* reflection problem of the type

$$q(\cdot) = q(0) + \int_0^{\cdot} \theta(t)\, du + \int_0^{\cdot} \left[I - P^T(t)\right] dy(t).$$

Similarly, the diffusion limit would be a solution of the time-dependent reflection problem

$$V(\cdot) = X(\cdot) + \int_0^{\cdot} \left[I - P^T(t)\right] dY(t),$$

$$X(\cdot) = V(0) + \left[I - I^-\right] \int_0^{\cdot} \left[I - P^T(t) I^-\right]^{-1} \times \left(-f_P^T(t)\, dy(t) + \Sigma^{1/2}(t)\, dW(t)\right).$$

We can thus deduce results for time-dependent networks from the corresponding results for state-dependent networks by introducing an auxiliary coordinate (additional station) in the manner outlined above. (The technique is similar to the way a differential equation is reduced to an autonomous equation.) However, this approach is restricted because it requires unnecessarily strong assumptions on rates and routing (in particular, Lipschitz continuity). We can, alternatively, pursue a direct approach that covers networks with discontinuous parameters. Such features are important in applications and are currently under study [56].

APPENDIX A

Projected differential equations. In what follows, we introduce state-dependent oblique projections. This notion is used in Section 5 to help characterize fluid limits. We begin with the following lemma:

LEMMA A.1. *Fix an arbitrary $\chi \in \mathcal{S} = \mathbb{R}_+^K$. For any $\xi \in \mathbb{R}^K$, there exists a unique pair $(z, v) \in \mathbb{R}^K \times \mathbb{R}^K$ such that*

$$\begin{aligned} &z = \xi + \left[I - P^T(\chi)\right] v, \\ &z \in T_{\mathcal{S}}(\chi), \qquad -v \in N_{\mathcal{S}}(\chi), \\ &z^T \cdot v = 0, \end{aligned} \tag{A.1}$$

where $T_{\mathcal{S}}(\chi)$ and $N_{\mathcal{S}}(\chi)$ are, respectively, the tangent and normal cones to $\mathcal{S}$ at χ.

REMARK. In the mathematical programming literature, (A.1) is known as the linear complementarity problem over cones (see, e.g., [19], page 31). The lemma can be derived from general results on the linear complementarity problem. However, to gain insight into the nature of (A.1), we provide a simple independent proof by relating the linear complementarity problem

(A.1) to the oblique reflection problem; that is, to the dynamic complementarity problem (see Remark B.1).

PROOF OF LEMMA A.1. Recall the expressions for $T_{\mathscr{S}}(\chi)$ and $N_{\mathscr{S}}(\chi)$ from (5.1). Pick an arbitrary $t_0 > 0$. For a given χ and ξ, introduce the piecewise linear function x,

$$x(t) = \chi + \xi(t - t_0)^+, \qquad t \geq 0. \tag{A.2}$$

By Theorem 1 in Harrison and Reiman [31], there exists a unique pair (q, y) of continuous functions satisfying

$$\begin{aligned} &q(t) = x(t) + \big[I - P^T(\chi)\big]y(t) \geq 0, \qquad t \geq 0, \\ &y \text{ is nondecreasing in each coordinate, } y(0) = 0, \\ &\int_0^\infty 1^T\big[q(t) > 0\big]\, dy(t) = 0. \end{aligned} \tag{A.3}$$

Moreover, for x given by (A.2), q and y are piecewise linear (see the proof of Theorem 5.2 in [14]). In particular, there exist $t_1 \in (t_0, \infty)$ and $b, c \in \mathbb{R}^K$ such that

$$q(t) = \chi + b(t - t_0)^+, \qquad y(t) = c(t - t_0)^+, \qquad t \in [0, t_1]. \tag{A.4}$$

Substituting (A.2) and (A.4) into (A.3) shows that

$$\begin{aligned} &b = \xi + \big[I - P^T(\chi)\big]c, \\ &b_i \geq 0 \quad \text{whenever } \chi_i = 0, \\ &c \in \mathbb{R}_+^K \quad \text{and} \quad c_i = 0 \quad \text{whenever } \chi_i > 0, \\ &b^T \cdot c = 0. \end{aligned} \tag{A.5}$$

Finally, let $z = b$ and $v = c$. From (A.5) it follows that these z and v satisfy (A.1), and thus existence of the solution to (A.1) is established. The uniqueness for (A.1) can be derived from uniqueness of the solution to (A.3) by applying the foregoing arguments in reverse order. □

REMARK. Roughly speaking, our proof illustrates that two piecewise linear functions satisfy the dynamic complementarity problem if and only if their slopes satisfy the linear complementarity problem over cones. (We can establish a similar relation for absolutely continuous functions, as well as for the jumps of step functions.)

Lemma A.1 supports the following definition:

DEFINITION A.1. Fix an arbitrary $\chi \in \mathscr{S}$ and $\xi \in \mathbb{R}^K$. Call the vector z satisfying (A.1) the state-dependent oblique projection of ξ onto $T_{\mathscr{S}}(\chi)$ with respect to $[I - P^T(\chi)]$ and denote it by

$$\Pi^{\mathscr{F}(\chi)}\{\xi\},$$

where $\mathscr{F}(\chi) = (T_{\mathscr{S}}(\chi), P(\chi))$.

APPENDIX B

Time-dependent reflection problems. We outline some new results on time-dependent reflection problems. Proofs are given in [57].

B.1. *Formulation of the problem and main properties.* Fix an arbitrary $T > 0$. Let J be a given subset of $\{1, 2, \ldots, K\}$ and

$$\Gamma_J = \{\xi \in \mathbb{R}^K\colon\ \xi_k \geq 0 \text{ for all } k \in J\}. \tag{B.1}$$

Throughout this Appendix, denote $\Gamma \triangleq \Gamma_J$, for simplicity of notation.

Let $P(\cdot) = [\,p_{jk}(\cdot)]_{j,k=1}^K$ be a given nonnegative matrix-valued function, $P\colon [0,T] \to \mathbb{R}_+^{K\times K}$, which is RCLL. Assume that P has the following properties: *first*, $P(t)$ is substochastic for every $t \in [0,T]$,

$$\sum_{k=1}^K p_{jk}(\cdot) \leq 1, \qquad j = 1, \ldots, K; \tag{B.2}$$

second, the spectral radii $r(P(\cdot))$ satisfy

$$\sup_{t\in[0,T]} r(P(t)) < 1. \tag{B.3}$$

DEFINITION B.1 (Time-dependent reflection problem). Let $\xi \in D_0^K[0,T]$. Then $(\phi, \psi) \in D^{2K}[0,T]$ is a solution to the time-dependent reflection problem for ξ (with respect to Γ and R) if

$$\begin{cases} \phi(t) = \xi(t) + \psi(t), \\ \phi(t) \in \Gamma, \qquad t \in [0,T]; \\ \text{there exists } y \in D^K[0,T] \text{ such that } \psi(\cdot) = \int_0^{\cdot} R(u)\,dy(u), \\ y \text{ is nondecreasing in each coordinate, } y(0) = 0,\ y_k \equiv 0 \\ \qquad \text{for all } k \notin J, \\ \int_0^T 1\{\phi_k(t) > 0\}\,dy_k(t) = 0, \qquad k \in J, \end{cases} \tag{B.4}$$

where $R(\cdot) \triangleq [\,I - P^T(\cdot)]$.

REMARK B.1. The special case when $R \equiv$ const is known as the *oblique reflection problem* [31] or the *dynamic complementarity problem* [54]. For the geometric interpretation of R as the matrix of directions of reflection, see Remark 4.9.

The following theorem plays a pivotal role in the proof of our FLLN and FCLT.

THEOREM B.1 (Existence, uniqueness and Lipschitz property). *Let* $\Gamma = \Gamma_J$ *be defined by* (B.1) *and let* P *satisfy* (B.2) *and* (B.3). *Assume also that* P *is*

absolutely continuous and

$$\|\dot{P}\|_T = L_T^d < \infty. \tag{B.5}$$

Then for each $\xi \in D_0^K[0,T]$ *there exists a unique pair* $(\phi, \psi) \in D^{2K}[0,T]$ *such that* (B.4) *is satisfied.*

Furthermore, let $\xi^1, \xi^2 \in D_0^K[0,T]$, *and let* (ϕ^1, ψ^1) *and* (ϕ^2, ψ^2) *be the solutions to* (B.4) *for* ξ^1 *and* ξ^2, *respectively. Then there exists* $L < \infty$ *such that*

$$\|\phi^1 - \phi^2\|_T \le L\|\xi^1 - \xi^2\|_T, \qquad \|\psi^1 - \psi^2\|_T \le L\|\xi^1 - \xi^2\|_T.$$

Moreover, this L depends solely on K, T, L_T^d *and* L_T^∞, *where*

$$L_T^\infty \triangleq \sup_{t\in[0,T]} \left|\left[I - P^T(t)\right]^{-1}\right|_\infty$$

is finite, by (C.2). *Finally, the functions* ϕ *and* ψ *are nonanticipating with respect to the data* ξ.

REMARK. Our proofs of existence and uniqueness are based on fixed-point theorems. Proof of the Lipschitz property develops ideas of Dupuis and Ishii [23]. This theorem can be generalized to the case when P is of bounded variation. However, the above partial results, covering absolutely continuous P, suffice for our purposes.

REMARK B.2. Based on Theorem B.1, we introduce two well-defined *time-dependent reflection operators* $\mathbf{\Phi}_\Gamma^{\mathbf{R}}$ and $\mathbf{\Psi}_\Gamma^{\mathbf{R}}$ by

$$\mathbf{\Phi}_\Gamma^{\mathbf{R}}(\xi) = \phi, \qquad \mathbf{\Psi}_\Gamma^{\mathbf{R}}(\xi) = \psi. \tag{B.6}$$

Both $\mathbf{\Phi}_\Gamma^{\mathbf{R}}$ and $\mathbf{\Psi}_\Gamma^{\mathbf{R}}$ are Lipschitz in the uniform metric and

$$\|\mathbf{\Phi}_\Gamma^{\mathbf{R}}(\xi)\|_T \le L\|\xi\|_T, \qquad \|\mathbf{\Psi}_\Gamma^{\mathbf{R}}(\xi)\|_T \le L\|\xi\|_T.$$

It is natural to introduce the notion of *state-dependent* reflection operators. These are associated with a corresponding state-dependent reflection problem. Here P and ξ may depend on the state ϕ. In particular, the reflection term ψ in (B.4) is given by

$$\psi(\cdot) = \int_0^{\cdot} \left[I - P^T(\phi(u))\right] dy(u).$$

We have no direct use of this concepts, but an example is (4.7). See also Remark 4.9.

PROPOSITION B.1. *Let* Γ *and P satisfy the conditions of Theorem* B.1. *If* ξ *is continuous* (*absolutely continuous*), *with* $\xi_k(0) \ge 0$, $k \in J$, *then* $\mathbf{\Phi}_\Gamma^{\mathbf{R}}$, $\mathbf{\Psi}_\Gamma^{\mathbf{R}}$ *and y in* (B.4) *are continuous* (*absolutely continuous*) *as well.*

PROPOSITION B.2. *Let* (ϕ, ψ) *be a solution to* (B.4) *for a given continuous* ξ *with respect to some* Γ *and P. Fix an arbitrary* $\tau \in (0,T)$. *Then* (ϕ^*, ψ^*),

given by

$$\phi^*(t) = \phi(\tau + t), \qquad \psi^*(t) = \psi(\tau + t) - \psi(\tau), \qquad t \in [0, T - \tau],$$

is a solution to (B.4) *for* $\xi^*(t) = \phi(\tau) + \chi(\tau + t) - \chi(\tau)$, *with respect to* Γ *and* $P^*(t) = P(\tau + t)$, *where* $t \in [0, T - \tau]$.

B.2. *Derivatives of time-dependent reflection operators*. Our goal here is to investigate the convergence of the sequence $\{V^n\}$ given by

$$V^n = \sqrt{n}\left[\mathbf{\Phi}^{\mathbf{R}}_{\mathscr{S}}\left(x + \frac{1}{\sqrt{n}}X\right) - \mathbf{\Phi}^{\mathbf{R}}_{\mathscr{S}}(x)\right], \qquad n = 1, 2, \ldots, \tag{B.7}$$

where $\mathscr{S} \triangleq \mathbb{R}^K_+$. The limit of $\{V^n\}$ can be interpreted as some form of a directional derivative of $\mathbf{\Phi}^{\mathbf{R}}_{\mathscr{S}}$, at the point x in the direction of X. The corresponding one-dimensional problem is treated by Mandelbaum and Massey [55].

In this paper we consider only cases when $\{V^n\}$ converges u.o.c. to a continuous limit. Treating discontinuous limits requires Skorokhod's M_1-topology. This issue was partially treated in [65] and will be addressed in a future generalization of [57]. For simplicity of presentation, we restricted ourselves to the smallest classes of functions x, X and P, which are sufficient for our applications. Specifically, we impose on these functions the following assumptions:

ASSUMPTIONS C.

(C1) x is absolutely continuous, with $x(0) \geq 0$ and with Lipschitz derivative $\theta = \dot{x}$.

(C2) X is continuous, with $X_k(0) \geq 0$ whenever $x_k(0) = 0$.

(C3) P (hence R) is absolutely continuous, such that (B.2), (B.3) and (B.5) are satisfied.

Next, denote

$$q^n \triangleq \mathbf{\Phi}^{\mathbf{R}}_{\mathscr{S}}\left(x + \frac{1}{\sqrt{n}}X\right), \qquad q \triangleq \mathbf{\Phi}^{\mathbf{R}}_{\mathscr{S}}(x)$$

and let y^n and y be the corresponding complementary functions given by (B.4). Note that, by Proposition B.1, q and y are absolutely continuous. Observe also that by the Lipschitz property of time-dependent reflection we have $\lim_{n \uparrow \infty} \|q^n - q\|_T = 0$. Within the context of our limit theorems, x and X play the following role: x is the driver of the fluid limit q, while X is the driver of the diffusion limit [cf. (7.19)]. In line with this interpretation, the limit V of $\{V^n\}$, given by Theorem B.2, plays the role of the diffusion limit [cf. (7.19)].

PROPOSITION B.3. *The function q is the unique solution to the projected DE* (*see Appendix A*)

$$\dot{q}(t) = \Pi^{\mathscr{F}(t)}\{\theta(t)\}, \qquad \mathscr{F}(t) = \big(T_{\mathscr{S}}(q(t)), P(t)\big), \qquad t \geq 0,$$

with the initial condition $x(0)$.

The proof is similar to that of Theorem 5.2 and therefore is omitted.

This proposition implies, in particular, that the function m given by

$$m(t) = \left[I - P^T(t)\right]^{-1}\left(\Pi^{\mathcal{F}(t)}\{\theta(t)\} - \theta(t)\right)$$

satisfies the following relation $m(t) = \dot{y}(t)$, for almost every $t > 0$.

For each moment t, define now the sets $J^+(t)$, $J^-(t)$ and $J^0(t)$ by

$$J^+(t) = \{j\colon q_j(t) > 0\},$$
$$J^-(t) = \{j\colon q_j(t) = 0,\ m_j(t) > 0\},$$
$$J^0(t) = \{j\colon q_j(t) = 0,\ m_j(t) = 0\}.$$

In our limit theorems, $J^+(t)$, $J^-(t)$ and $J^0(t)$ act as the sets of overloaded, underloaded and critically loaded stations [cf. (6.2)]. Finally, we set $X(0-) = 0$.

The following main theorem provides a deterministic framework for our FCLT. In particular, the theorem demonstrates that the diffusion limits for overloaded, critically loaded and underloaded stations are processes without reflection, with reflection and the zero process, respectively.

THEOREM B.2. *Assume the following statements*:

(i) *Assumptions C are satisfied.*
(ii) $J^+(\cdot)$, $J^-(\cdot)$ *and* $J^0(\cdot)$ *are constant during* $[0, T]$.
(iii) $X_k(0) = 0$ *for all* $k \in J^-$.

Then the sequence $\{V^n\}$ *converges uniformly over* $[0, T]$ *to a function* V. *This* V *is the unique solution to the time-dependent reflection problem*

$$V = \mathbf{\Phi}_{\Gamma}^{\mathbf{R}}\left(\left[I - I^-\right]\int_0^{\cdot} \tilde{R}^{-1}(t)\, dX(t)\right),$$
$$\tilde{R} = I - P^T I^-, \qquad \Gamma = \left\{\xi \in \mathbb{R}^K\colon \xi_k \geq 0,\ \forall k \in J^- \cup J^0\right\}.$$

REMARK B.3. Note that $V(0) = X(0)$, as it must be according to (7). Indeed, similarly to [15], write

$$I^- = \begin{bmatrix} 0 & 0 \\ 0 & I_N \end{bmatrix}, \qquad P = \begin{bmatrix} P_B & P_{BN} \\ P_{NB} & P_N \end{bmatrix}, \qquad X(0) = \begin{bmatrix} X_B \\ 0 \end{bmatrix}.$$

Simple calculations yield

$$\left[I - I^-\right]\left[I - P^T I^-\right]^{-1} = \begin{bmatrix} I_B & P_{NB}^T\left[I_N - P_N^T\right]^{-1} \\ 0 & 0 \end{bmatrix}.$$

Hence $V(0) = [I - I^-][I - P^T(0)I^-]^{-1}X(0) = X(0)$.

APPENDIX C

Properties of the routing matrix P. Let $P(\cdot) = [p_{jk}(\cdot)]_{j,k=1}^{K}$ be a given nonnegative matrix-valued function, $P\colon \mathfrak{V} \to \mathbb{R}_+^{K\times K}$. Here $\mathfrak{V}$ denotes any one of $[0,T]$, $[0,\infty)$ or $\mathbb{R}_+^K$. Suppose further that the following conditions are satisfied:

$$\sum_{k=1}^{K} p_{jk}(\cdot) \le 1, \qquad j = 1,\dots,K, \quad \text{and} \quad \sup_{\xi\in\mathfrak{V}} r(P(\xi)) < 1.$$

By the Perron–Frobenius theorem [8], these conditions imply, in a straightforwared manner, that the matrix $[I - P^T(\xi)]$ is invertible for every $\xi \in \mathfrak{V}$, and

$$\inf_{\xi\in\mathfrak{V}} \det\left[I - P(\xi)\right] > 0, \tag{C.1}$$

$$\sup_{\xi\in\mathfrak{V}} \left|\left[I - P^T(\xi)\right]^{-1}\right| < \infty, \tag{C.2}$$

$$\max_{1\le i\le K} \sup_{\xi\in\mathfrak{V}} p_{ii}(\xi) < 1. \tag{C.3}$$

APPENDIX D

Notation.

Sets.

$\mathscr{Z}_+$ and $\mathbb{R}_+$	the sets of nonnegative integer and real numbers
$\mathbb{R}^K$	the K-dimensional Euclidean space
$\mathscr{S} \triangleq \mathbb{R}_+^K$	$\{\xi \in \mathbb{R}^K\colon \xi_k \ge 0, \text{ for all } k = 1,\dots,K\}$
$\mathscr{S}^0$	the interior of $\mathscr{S}$
$\mathbb{R}_-^K$	$\{\xi \in \mathbb{R}^K\colon \xi_k \le 0, \text{ for all } k = 1,\dots,K\}$
$\mathbb{R}_+^{K\times K}$	the set of $K\times K$-dimensional matrices with nonnegative elements
$\mathbb{R}^{K\times K\times K}$	the set of $K\times K\times K$ arrays with real elements
$\mathscr{E}[0,1]$	the set of all open subintervals of $[0,1]$

Vector and matrices.

a^T	transpose of a vector or a matrix
$\det[P]$ and $r(P)$	the determinant and the spectral radius of a matrix P
I and δ_{jk}	the identity matrix and Kronecker's symbol
$\operatorname{diag}\{a\}$, $a \in \mathbb{R}^K$	the matrix $\operatorname{diag}\{a_1,\dots,a_K\}$
$\lvert a\rvert$	Euclidean norm of a vector a
$\lvert P\rvert$	operator norm of a matrix P with respect to Euclidean vector norm

$\|a\|_1$ and $\|a\|_\infty$, $a \in \mathbb{R}^K$	$\sum_{k=1}^K \|a_k\|$ and $\max_k \|a_k\|$
$\|P\|_\infty$, $P \in \mathbb{R}^{K\times K}$	$\max_k \sum_{j=1}^K \|p_{kj}\|$
$\mathbf{\Theta}\odot a$, $\mathbf{\Theta} \in \mathbb{R}^{K\times K\times K}$, $a \in \mathbb{R}^K$	$[\sum_{i=1}^K \mathbf{\Theta}_{ijk} a_i]_{j,k=1}^K$

Derivatives of vector and matrix functions.

$\partial\theta(\xi)$, $\theta\colon \mathbb{R}^K \to \mathbb{R}^K$	$[(\partial\theta_j(\xi))/\partial\xi_k]_{j,k=1}^K$
$\partial P(\xi)$, $P\colon \mathbb{R}^K \to \mathbb{R}^{K\times K}$	$[(\partial P_{jk}(\xi))/\partial\xi_i]_{i,j,k=1}^K$

Function spaces.

$D^K[0,T]$ and $C^K[0,T]$	the set of RCLL and continuous $\mathbb{R}^K$-valued functions on $[0,T]$
$D_0^K[0,T]$ and $C_0^K[0,T]$	$\{\xi \in D^K[0,T]\colon \xi_k(0) \geq 0,\ k \in J\}$ and $\{\xi \in C^K[0,T]\colon \xi_k(0) \geq 0,\ k \in J\}$
$\|a\|_T$	$\sup_{0\leq t\leq T}\|a(t)\|$, where a is a vector or a matrix endowed with uniform topology

Convergence.

u.o.c.	uniformly on compact
$\to_d$ and $\to_p$, $\mathbf{P}-\lim$	convergence in distribution and in probability

Stochastic processes.

RCLL	right-continuous with left limits
$\langle M \rangle$	the predictable quadratic variation of a martingale M
$\langle M^1, M^2 \rangle$	the predictable quadratic covariation of martingales M^1 and M^2

Miscellaneous.

$\{a^n\}$	sequence a^n, $n = 1, 2, \ldots$
$\vee$ and $\wedge$	maximum and minimum
$a^+ = a \vee 0$ and $a^- = -(a \wedge 0)$	the positive and negative parts of a
$\bar{f}(t) = \sup_{0\leq s\leq t} f_s$	the upper envelope of f
$1\{S\}$	indicator function of a set S
$1\{\xi > 0\}$ and $1\{\xi = 0\}$; $\xi \in \mathbb{R}^K$	$(1\{\xi_1 > 0\}, \ldots, 1\{\xi_K > 0\})^T$ and $(1\{\xi_1 = 0\}, \ldots, 1\{\xi_K = 0\})^T$
$I\{\xi > 0\}$ and $I\{\xi = 0\}$; $\xi \in \mathbb{R}^K$	$\mathrm{diag}\{1\{\xi > 0\}\}$ and $\mathrm{diag}\{1\{\xi = 0\}\}$
$\mathscr{J}_0(\xi)$, $\xi \in \mathbb{R}^K$	$\{k\colon \xi_k = 0\}$
J^+, J^-, J^0	see (6.2)
I^+	$\mathrm{diag}\{1\{1 \in J^+\}, \ldots, 1\{K \in J^+\}\}$
I^-, I^0	$\mathrm{diag}\{1\{1 \in J^-\}, \ldots, 1\{K \in J^-\}\}$, $\mathrm{diag}\{1\{1 \in J^0\}, \ldots, 1\{K \in J^0\}\}$

Conventions. Vectors are understood to be columns. For a convex set $\Gamma \subseteq \mathbb{R}^K$ and $\chi \in \Gamma$, $N_\Gamma(\chi)$ and $T_\Gamma(\chi)$ denote, respectively, the normal and tangent cones at χ:

$$N_\Gamma(\chi) = \{\zeta \in \mathbb{R}^K\colon \zeta^T \cdot (\xi - \chi) \le 0 \text{ for all } \xi \in \Gamma\},$$

$$T_\Gamma(\chi) = \{\zeta \in \mathbb{R}^K\colon \zeta^T \cdot \xi \le 0 \text{ for all } \xi \in N_\Gamma(\chi)\}.$$

A vector- or matrix-valued function f is locally Lipschitz if for every compact set $\mathscr{K} \subset \mathbb{R}^K$, there exists a constant $L^{\mathscr{K}}$ such that

$$|f(\xi^1) - f(\xi^2)|_\infty \le L^{\mathscr{K}} |\xi^1 - \xi^2|, \qquad \xi^1, \xi^2 \in \mathbb{R}^K;$$

f is globally Lipschitz if $L^{\mathscr{K}}$ can be chosen independently of K. Integrals $\int_0^t$ stand for $\int_{[0,t]}$.

Acknowledgments. A. M. was supported by the Fund for Promotion of Research at the Technion. G. P. acknowledges the support of the Technion, the Mrs. Kennedy Leigh Fund and the Council for Higher Education. This paper originated as the Technion Ph.D. thesis of G. P., supervised by A. M. Part of this research was carried out while G. P. was a postdoctoral fellow at Tel-Aviv University; the hospitality of the Department of Statistics and Operations Research is greatly appreciated. The authors are grateful to the Editor and an anonymous referee for helpful and constructive comments.

REFERENCES

[1] ABATE, J. and WHITT, W. (1987). Transient behavior of regulated Brownian motion, I: starting at the origin. *Adv. in Appl. Probab.* **19** 560–598.

[2] ABATE, J. and WHITT, W. (1987). Transient behavior of regulated Brownian motion, II: non-zero initial condition. *Adv. in Appl. Probab.* **19** 599–631.

[3] ANICK, D., MITRA, D. and SONDHI, M. M. (1982). Stochastic theory of a data-handling system with multiple sources. *Bell System Tech. J.* **61** 1871–1894.

[4] ANISIMOV, V. V. (1997). Switching processes: asymptotic theory and applications. In *New Trends in Probability and Statistics* (A. N. Shiryayev et al., eds.). To appear.

[5] ANULOVA, S. V. (1990). Functional limit theorems for network of queues. In *Proc. IFAC Congress*, Abstract.

[6] AUBIN, J. P. and CELLINA, A. (1984). *Differential Inclusions*. Springer, New York.

[7] BARBOUR, A. D. (1974). On a functional central limit theorems for Markov population processes. *Adv. in Appl. Probab.* **6** 21–39.

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

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

[10] BOUCHERIE, R. J. and VAN DIJK, N. M. (1993). A generalization of Norton's theorem for queueing networks. *Queueing Systems Theory Appl.* **13** 251–289.

[11] BREMAUD, P. (1981). *Point Processes and Queues*: *Martingale Dynamics*. Springer, Berlin.

[12] BUZACOTT, J. A. and YAO, D. D. (1986). On queueing network models of flexible manufacturing systems. *Queueing Systems Theory Appl.* **1** 5–27.

[13] CHEN, H. (1990). Generalized regulated mapping, fluid and diffusion limits. Unpublished notes.

[14] CHEN, H. and MANDELBAUM, A. (1991). Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.* **16** 408–446.

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

[16] CLARKE, F. H. (1983). *Optimization and Nonsmooth Analysis*. Wiley, New York.
[17] COFFMAN, E. G., JR., PUHALSKII, A. A., REIMAN, M. I. and WRIGHT, P. (1994). Processor shared buffers with reneging. *Performance Evaluation* **19** 25–46.
[18] COFFMAN, E. G., JR., PUHALSKII, A. A. and REIMAN, M. I. (1991). Storage limited queues in heavy traffic. *Probab. Engrg. Inform. Sci.* **5** 499–522.
[19] COTTLE, R. W., PANG, J. S. and STONE, R. E. (1992). *The Linear Complementarity Problem*. Academic Press, New York.
[20] DAVIS, G. A. and NIHAN, N. L. (1993). Large population approximations of a general stochastic traffic assignment model. *Oper. Res.* **41** 169–178.
[21] DUPUIS, P. and ISHII, H. (1990). On oblique derivative problems for fully nonlinear second-order elliptic partial differential equations on nonsmooth domains. *Nonlinear Anal.* **15**(12) 1123–1138.
[22] DUPUIS, P. and ISHII, H. (1991). On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications. *Stochastics Stochastics Rep.* **35** 31–62.
[23] DUPUIS, P. and ISHII, H. (1993). SDEs with oblique reflection on nonsmooth domains. *Ann. Probab.* **21** 554–580.
[24] ELLIOTT, R. J. (1982). *Stochastic Calculus and Applications*. Springer, New York.
[25] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Process: Characterization and Convergence*. Wiley, New York.
[26] FILIPPOV, A. F. (1988). *Differential Equations with Discontinuous Right Hand Sides*. Kluwer Academic, Dordrecht.
[27] GIORNO, V., NEGRI, C. and NOBILE, A. (1985). A solvable model for a finite-capacity system. *J. Appl. Probab.* **22** 903–911.
[28] HALE, J. (1969). *Ordinary Differential Equations*. Wiley, New York.
[29] HALFIN, S. and WHITT, W. (1981). Heavy-traffic limits theorem for queues with many exponential servers. *Oper. Res.* **29** 567–588.
[30] HARRISON, J. M. (1985). *Brownian Motion and Stochastic Flow Systems*. Wiley, New York.
[31] HARRISON, J. M. and REIMAN, M. I. (1981). Reflected brownian motion on an orthant. *Ann. Probab.* **9**(2) 302–308.
[32] HARRISON, J. M. and SHEPP, L. A. (1984). A tandem storage system and its diffusion limit. *Stochastic Process. Appl.* **16** 257–274.
[33] HEFFES, H. (1982). Moment formulae for a class of mixed multi-job-type queueing networks. *Bell System Tech. J.* **61** 709–745.
[34] IGLEHART, D. L. (1965). Limit diffusion approximations for the many-server queue and repairman problem. *J. Appl. Probab.* **2** 429–441.
[35] IGLEHART, D. L. and LEMOINE, A. J. (1973). Approximations for the repairman problem with two repair facilities, I: no spares. *Adv. in Appl. Probab.* **5** 595–613.
[36] ISHAM, V. (1993). Stochastic models for epidemics with special reference to AIDS. *Ann. Appl. Probab.* **3**(1) 1–27.
[37] JACKSON, R. J. (1963). Jobshop-like queueing systems. *Management Sci.* **10** 131–142.
[38] JACOD, J. and SHIRYAYEV, A. N. (1987). *Limit Theorems for Stochastic Processes*. Springer, Berlin.
[39] JENNINGS, O. B., MANDELBAUM, A., MASSEY, W. A. and WHITT, W. (1996). Server staffing to meet time varying demand. *Management Science* **42** 1383–1394.
[40] KARATZAS, I. and SHREVE, S. E. (1988). *Brownian Motion and Stochastic Calculus*. Springer, New York.
[41] KASPI, H. and MANDELBAUM, A. (1992). Regenerative closed queueing networks. *Stochastics Stochastics Rep.* **39** 239–258.
[42] KLOEDEN, P. and PLATEN, E. (1992). *Numerical Solutions of Stochastic Differential Equations*. Springer, Berlin.
[43] KNESSL, C. and MORRISON, J. A. (1991). Heavy-traffic analysis of a data-handling system with many sources. *SIAM J. Appl. Math.* **51** 187–213.
[44] KOGAN, Y. A. and LIPTSER, R. S. (1993). Limit non-stationary behavior of large closed queueing networks with bottlenecks. *Queueing Systems Theory Appl.* **14** 33–55.

[45] KOGAN, Y. A., LIPTSER, R. S. and SMORODINSKII, A. V. (1986). Gaussian diffusion approximation of closed Markov models of computer networks. *Problems Inform. Transmission* **22** 38–51.
[46] KRICHAGINA, E. V. (1992). Asymptotic analysis of queueing networks (martingale approach). *Stochastics Stochastics Rep.* **40** 43–76.
[47] KRZESINSKI, A. E. (1987). Multiclass queueing networks with state-dependent routing. *Performance Evaluation* **7** 125–143.
[48] KURTZ, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Probab.* **7** 49–58.
[49] KURTZ, T. G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.* **8** 344–356.
[50] KURTZ, T. G. (1978). Strong approximation theorems for density dependent Markov chains. *Stochastic Process. Appl.* **6** 223–240.
[51] KURTZ, T. G. (1980). Representation of Markov processes as multiparameter time changes. *Ann. Probab.* **8** 682–715.
[52] KUSHNER, H. J. and MARTINS, L. F. (1993). Heavy traffic analysis of a data transmission systems with many independent sources. *SIAM J. Appl. Math.* **53** 1095–1122.
[53] LIPTSER, R. S. and SHIRYAYEV, A. N. (1989). *Theory of Martingales*. Kluwer Academic, Dordrecht.
[54] MANDELBAUM, A. (1989). The dynamic complementarity problem. Preprint.
[55] MANDELBAUM, A. and MASSEY, W. A. (1995). Strong approximations for time-dependent queues. *Math. Oper. Res.* **20** 33–64.
[56] MANDELBAUM, A., MASSEY, W. A. and PATS, G. (1997). Approximations for time-dependent networks. Unpublished manuscript.
[57] MANDELBAUM, A., MASSEY, W. A. and PATS, G. (1995). Time-dependent reflection problem. Technical Report, Technion.
[58] MANDELBAUM, A., MASSEY, W. A. and REIMAN, M. I. (1997). Strong approximations for Markovian service networks. Preprint.
[59] MANDELBAUM, A. and PATS, G. (1995). State-dependent queues: approximations and applications. In *IMA Volumes in Mathematics and Its Applications* (F. Kelly and R. J. Williams, eds.) **71** 239–282. Springer, Berlin.
[60] MASSEY, W. A. and WHITT, W. (1993). Networks of infinite-server queues with non-stationary Poisson input. *Queueing Systems Theory Appl.* **13** 183–250.
[61] MITRA, D. (1988). Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Adv. in Appl. Probab.* **20** 646–676.
[62] MITRA, D. and MCKENNA, J. (1986). Asymptotic expansions for closed Markovian networks with state-dependent service rates. *J. Assoc. Comput. Mach.* **33** 568–592.
[63] MITRANI, I. and PUHALSKII, A. A. (1993). Limiting results for multiprocessor systems with breakdowns and repairs. *Queueing Systems Theory Appl.* **14** 293–311.
[64] PATS, G. (1997). State-dependent queueing networks. Part II: approximations with discontinuous diffusion limits. Unpublished manuscript.
[65] PATS, G. (1994). State-dependent queueing networks: approximations and applications. Ph.D. thesis, Faculty of Industrial Engineering and Management, Technion.
[66] PEGDEN, C., SHANNON, R. and SADOWSKI, R. (1995). *Introduction to Simulation Using SIMAN*, 2nd ed. McGraw-Hill, New York.
[67] POMAREDE, J. L. (1976). A unified approach via graphs to Skorokhod's topologies on the function space. Ph.D. dissertation, Dept. Statistics, Yale Univ.
[68] PRISGROVE, L. A. (1987). Closed queueing networks with multiple servers: transient and steady-state approximations. Technical Report 20, Dept. Operations Research, Stanford Univ.
[69] PROTTER, P. (1990). *Stochastic Integration and Differential Equations*. Springer, Berlin.
[70] SERFOZO, R. F. (1989). Markovian network processes: congestion-dependent routing and processing. *Queueing Systems Theory Appl.* **5** 5–36.
[71] SERFOZO, R. F. (1993). Queueing networks with dependent nodes and concurrent movements. *Queueing Systems Theory Appl.* **13** 143–182.

[72] SLOMINSKI, L. (1989). Stability of strong solutions of stochastic differential equations. *Stochastic Process. Appl.* **31** 173–202.
[73] STROOK, D. W. and VARADHAN, S. R. S. (1979). *Multidimensional Diffusion Processes*. Springer, Berlin.
[74] SYSKI, R. (1986). *Introduction to Congestion Theory in Telephone Systems*, 2nd ed. North-Holland, Amsterdam.
[75] TOWSLEY, D. F. (1980). Queueing network models with state-dependent routing. *J. Assoc. Comput. Mach.* **27** 323–337.
[76] VANDERGRAFT, A. J. (1983). A fluid model of networks of queues. *Management Sci.* **29** 1198–1208.
[77] VAN DIJK, N. M. (1993). *Queueing Networks and Product Forms*. Wiley, New York.
[78] WHITE, J. A., SCHMIDT, J. W. and BENNETT, G. K. (1974). *Analysis of Queueing Systems*. Academic Press, New York.
[79] WHITT, W. (1984). Open and closed models for networks of queues. *Bell System Tech. J.* **63** 1911–1979.
[80] WHITT, W. (1992). Understanding the efficiency of multi-server service systems. *Management Sci.* **38** 708–723.
[81] WORTHINGTON, D. J. (1987). Queueing models for hospital waiting lists. *J. Oper. Res. Soc.* **38** 413–422.
[82] YAMADA, K. (1986). Multi-dimensional Bessel processes as heavy traffic limits of certain tandem queues. *Stochastic Process. Appl.* **23** 35–56.
[83] YAMADA, K. (1993). Diffusion approximations for open state-dependent queueing networks under heavy traffic situation. Technical Report, Inst. Information Science and Electronics, Univ. Tsukuba, Japan.
[84] YAO, D. D. and BUZACOTT, J. A. (1985). Modelling of a class of state-dependent routing in flexible manufacturing systems. *Ann. Oper. Res.* **3** 153–167.
[85] YAO, D. D. and BUZACOTT, J. A. (1986). Models of flexible manufacturing systems with limited local buffers. *International Journal of Production Research* **24** 107–117.

FACULTY OF INDUSTRIAL ENGINEERING AND MANAGEMENT
TECHNION–ISRAEL INSTITUTE OF TECHNOLOGY
TECHNION CITY, HAIFA 32000
ISRAEL
E-MAIL: avim@tx.technion.ac.il
gennadi@ie.technion.ac.il