FLUID AND DIFFUSION LIMITS FOR QUEUES IN SLOWLY CHANGING ENVIRONMENTS

G. L. CHOUDHURY
AT&T Laboratories
Room 1L-238
Holmdel, NJ 07730-3030

A. MANDELBAUM
Faculty of Industrial and
Management Engineering
The Technion, Haifa 32000
ISRAEL

M. I. REIMAN
Bell Labs
Lucent Technologies
Room 2C-117
Murray Hill, NJ 07974-0636

W. WHITT
AT&T Laboratories
Room 2C-178
Murray Hill, NJ 07974-0636

Stochastic Models 13 (1997) 121–146
(Received the first Marcel F. Neuts best paper award)

Key Words: queues in random environments, nearly-completely-decomposable Markov chains, Markov-modulated process, piecewise-stationary $M_t/G_t/1$ queue, stochastic fluid models, diffusion processes, fluid limit, heavy traffic.

ABSTRACT

We consider an infinite-capacity s-server queue in a finite-state random environment, where the traffic intensity exceeds 1 in some environment states and the environment states change slowly relative to arrivals and service completions. Queues grow in unstable environment states, so that it is useful to look at the system in the time scale of mean environment-state sojourn times. As the mean environment-state sojourn times grow, the queue-length and workload processes grow. However, with appropriate normalizations, these processes converge to fluid processes and diffusion processes. The diffusion process in a random environment is a refinement of the fluid process in a random environment. We show how the scaling in these limits can help explain numerical results for queues in slowly changing random environments. For that purpose, we apply recently developed numerical-transform-inversion algorithms for the

MAP/G/1 queue and the piecewise-stationary $M_t/G_t/1$ queue.

1. Introduction

An important feature of many complex systems is the existence of different behavior on different time scales. For example, in communication and computer systems the relevant time scale for users may be seconds, while the relevant time scale for system transactions may be milliseconds or microseconds.

In stochastic models, different time scales can be represented by nearly-completely-decomposable (NCD) Markov chains; e.g., see Simon and Ando [38], Courtois [12], Latouche [26], Philippe, Saad and Stewart [34], Chang and Nelson [3] and Latouche and Schweitzer [28]. In an NCD Markov chain, the state space can be decomposed into subsets such that the chain usually tends to move around within each subset and only rarely moves from one subset to another. If the local chain is ergodic, then the chain tends to reach a local steady-state within each subset before it leaves. Thus, in a longer time scale the chain tends to move from one steady-state regime to another, so that the long-run steady-state probability distribution is an average of the local steady-state distributions.

However, very different behavior can occur if local steady state cannot be reached in some of the subsets. Local steady state will not happen if the Markov chain is transient or null-recurrent within such a subset. This phenomenon arises naturally in an infinite-capacity queue in a slowly changing environment, where the queue is unstable in some environment states. The simplest example is perhaps the NCD Markov-modulated M/M/1 queue considered by Latouche and Schweitzer [28]. Their model is the M/M/1 queue

in a finite-state Markovian environment, where the environment states change very slowly and in some environment states the arrival rate is greater than the service rate. Latouche and Schweitzer study the steady-state distribution as the expected sojourn times in the environment states tend to infinity.

If the queue is unstable in some environment states, and if the environment states change slowly, then it is useful to focus on such systems in the time scale of expected sojourns in environment states. We show that, if we scale space and time appropriately, then the families of queue-length and workload processes converge to stochastic fluid processes (fluid processes in random environments) as the expected environment-state sojourn times go to infinity. After performing the scaling for the fluid limit, the scaled environment process tends to become fixed, while the arrival and service rates in each environment state are accelerated. (It is possible to think of the system with this scaling from the outset.) The acceleration of the arrival and service rates makes the law of large numbers applicable to the arrival and service processes, and thus to the queueing processes, within each environment state.

The first implication of this result is that it is natural to study such queueing systems via approximating stochastic fluid processes. To a large extent, this view is already taken when people directly introduce Markov-modulated fluid models to represent discrete queueing systems, e.g., see Anick, Mitra and Sondhi [1], Mitra [32], Stern and Elwalid [39], Elwalid and Mitra [14], Rogers [37], Kulkarni [24] and references cited there. Nevertheless, it is useful to make the connection between the models clear.

Once we see the connection between the fluid models and the associated queueing models, we realize that it may not actually be necessary to separately consider the fluid models. We show that any Markov-modulated fluid

model can be represented as a limit of MMPP/G/1 queues, where G can be any distribution, including exponential (M). The MMPP/M/1 model is especially tractable because its queue-length process is a quasi-birth-death (QBD) process. A new efficient algorithm for QBD process has been obtained by Latouche and Ramaswami [27]. Moreover, we show by example that the limit is approached sufficiently quickly that algorithms for the MMPP/G/1 queue can serve as algorithms for Markov-modulated fluid models. We expect that direct algorithms for stochastic fluid models will usually be more efficient than MMPP/G/1 queueing algorithms applied to queueing-model approximations of fluid models, but it is important to recognize that the set of all Markov-modulated fluid models can be regarded as being contained in the closure of the set of MMPP/M/1 queueing models.

The second implication of the fluid limits is that much of the queueing system behavior in this setting can be understood through the *scaling* in the fluid limit theorem: Remarkable statistical regularity in numerical results for the queueing systems can be seen when the scaling is done. We contend that the limit theorem is important, not only for generating approximations for intractable systems, but also to interpret the results for tractable systems. To show how well the scaling explains results for queues in slowly changing random environments, we exploit numerical-transform-inversion algorithms for solving the piecewise-stationary $\mathrm{M_t/G_t/1}$ queue in Choudhury, Lucantoni and Whitt [8] and the MAP/G/1 queue in Lucantoni [29], Lucantoni, Choudhury and Whitt [30], Choudhury and Whitt [10] and Choudhury, Lucantoni and Whitt [9].

We also gain insight by looking at how the limiting fluid process depends on the queueing model. It turns out that the behavior of the environment process

is critical, whereas the behavior of the queueing processes within environment states is not critical. This is easy to understand from the perspective of time scales. Changes in the environment occur in the important long time scale, whereas changes in the queue within an environment state occur in the less important short time scale. Thus, the limiting results still hold with the same fluid process limits for the more general infinite-capacity s-server G/G/s queue in the same Markovian environment (which requires some care in definition). It is significant that the G/G/s structure affects the limit only via the number of servers and the arrival and service rates; i.e., there is an *asymptotic insensitivity* to the arrival process and service-time sequences within environment states beyond their first moments. Moreover, in the limit the multiple-server feature tends to play a minor role. As in previous heavy-traffic limits, the system with s servers behaves the same as a single-server system with s times the service rate.

On the other hand, changes in the environment process can significantly affect the limit. The continuous-time Markov-chain (CTMC) environment process has exponential sojourn times in each state. After scaling, these sojourn times again have exponential distributions, but now with mean of order $O(1)$. The limiting fluid process in a random environment can be significantly different if the CTMC environment process is changed to a non-Markov process.

In some cases, especially with multiple servers, the queueing models are not easy to solve directly. Then the limiting stochastic fluid models should often be much easier to analyze. For example, consider the case of a G/G/s queue with $s > 1$ in a two-state alternating-renewal-process environment. The G/G/s model with $s > 1$ can be difficult even with only a single environment state. And there is not yet a solution to the G/G/s queue in a two-state

environment. However, the limiting stochastic fluid model with a two-state environment is quite nice. It has the structure of the GI/G/1 queue; see Kella and Whitt [22].

This paper is intended to contribute to the growing literature on fluid limits based on laws of large numbers, e.g., see Chen and Mandelbaum [5], Chen and Yao [7] and Section 7.5 of Kella and Whitt [23]. In fact, the fluid limit here can be regarded as a minor modification of previous heavy-traffic limit theorems in Iglehart and Whitt [18], [19]. If we condition on the environment process, then we have a sequence of fixed sojourns intervals in different environment states. We can then proceed inductively, establishing the fluid limit in each successive interval by using the standard heavy-traffic argument. The early papers [18], [19] considered only heavy-traffic diffusion processes limits, but it was known then that there existed corresponding laws of large numbers with deterministic fluid limits, even though the word "fluid" was not used; e.g., see Iglehart [17] and Whitt [40], [41], [42] for discussion. The interesting feature here is the random environment process.

Just as the central limit theorem provides a refinement to the law of large numbers, convergence to a diffusion process (Brownian motion) in a random environment provides a refinement to the fluid limits. Again, these diffusion limits follow by standard heavy-traffic arguments once we condition on the environment process. After we condition on the environment process, the queue can be regarded as time-dependent, and the heavy-traffic limiting behavior is as described for the time-dependent M/M/1 queue by Mandelbaum and Massey [31]. The behavior is somewhat simpler here, because our time-dependent model is piecewise stationary. Whenever the fluid process is positive, the diffusion process (Brownian motion) describes fluctuations in the

queue around the fluid process. When the fluid process is staying at 0, we need to look more carefully at the queue. If the queue is strictly stable ($\rho < 1$), then there is no diffusion or, equivalently, its variance parameter is 0. However, if $\rho = 1$ in the queue, then the diffusion approximation is reflected Brownian motion (RBM) with zero drift in this region. We avoid this case in the present paper by assuming that always $\rho > 1$ or $\rho < 1$. (We intend to address the case $\rho = 1$ in a subsequent paper.)

Since the diffusion approximation is a refinement to the fluid approximation, the diffusion approximation is more accurate. However, as shown for the MMPP/G/1 queue in Section 4, the stochastic fluid approximation itself can be quite accurate. Although the diffusion process in a random environment is just a Brownian motion (with variance depending on the environment), it is difficult to obtain its stationary distribution. An alternative diffusion approximation is RBM in a random environment, which has been considered by Asmussen [2] and Karandikar and Kulkarni [20].

There is a setting when the stochastic fluid model is especially appropriate, namely, when the diffusion limit coincides with the fluid limit. This will occur when our G/G/s queue in a random environment is essentially a D/D/s queue in a random environment. Bounds on the approximation error for the D/D/1 case can be constructed as in Chen and Yao [7]. Variants of D/D/1 queues in a random environment are often appropriate models for emerging high-speed communication networks and manufacturing systems. Thus there is especially strong motivation for the stochastic fluid models in some contexts, e.g., as in Anick et al. [1], Mitra [32], Stern and Elwalid [39] and Elwalid and Mitra [14]. Nevertheless, by virtue of the fluid limits, it is possible to represent these fluid models as limits of MMPP/M/1 queues and thus apply algorithms for the

queueing models to solve the fluid models. This may be counter to intuition until one realizes that Poisson processes with high rate behave essentially the same as deterministic processes.

Here is how the rest of this paper is organized. We begin in Section 2 by extending the NCD Markov-modulated M/M/1 queue considered by Latouche and Schweitzer [28] to the s-server analog and indicating what the fluid limits for the queue-length and workload processes should be. In Section 3 we extend the model to a general G/G/s model in a general stochastic (possibly non-Markovian) environment. We establish stochastic fluid limits for this much more general model by assuming a conditional strong law of large numbers for the interarrival times and service times given that the environment is in a particular state at a particular time. This general assumption covers many standard examples and seems natural to make as a direct assumption in applications. In Section 4 we consider a numerical example for the MMPP/G/1 queue with a fixed general service-time distribution and an MMPP arrival process with slowly changing environments. We show how the scaling in the limits help interpret the numerical results.

In Section 5 we discuss diffusion process limits that serve as refinements to the fluid limits in Section 3. Our key assumption is a conditional functional central limit theorem for the interarrival times and service times after an environment process transition. In Section 6 we consider a numerical example for the piecewise-stationary $\mathrm{M_t/G_t/1}$ queue. Finally, in Section 7 we give the proofs of the two theorems.

The limits we establish extend to queueing networks in slowly changing environments, drawing on Chen and Mandelbaum [5], in the spirit of Chen and Whitt [6], but we do not discuss networks here.

2. The NCD Markov-Modulated M/M/s Queue

We start by defining a family of Markovian queueing processes, which serve as concrete examples of the systems we consider; this model illustrates a particular case when the general conditions in Section 3 hold. Our model here is the multi-server generalization of the NCD Markov-modulated M/M/1 queue considered by Latouche and Schweitzer [28].

We introduce a *reference environment process*, which is an irreducible finite-state continuous-time Markov chain (CTMC) $X \equiv \{X(t) : t \geq 0\}$ with infinitesimal generator $Q \equiv (Q_{ij})$ of order m. As usual, the off-diagonal elements of Q are nonnegative and the row sums are zero. We make a *family of slowly changing environment processes* $X_\epsilon \equiv \{X_\epsilon(t) : t \geq 0\}$ by simply slowing down time:

$$X_\epsilon(t) = X(\epsilon t), \ t \geq 0 \ , \ \text{for } \epsilon > 0 \ . \tag{2.1}$$

It is easy to see that the CTMC X_ϵ has infinitesimal generator $Q_\epsilon \equiv \epsilon Q$.

For each $\epsilon > 0$, define a Markov-modulated M/M/s queue-length process $N_\epsilon \equiv \{N_\epsilon(t) : t \geq 0\}$ by letting N_ϵ evolve as an M/M/s queue with s_i servers, arrival rate λ_i and individual service rate μ_i whenever the environment process X_ϵ is in state i. We assume that s_i is finite for all i and that each M/M/s model has infinite capacity and the first-come first-served discipline. The Markov property implies that there are no end effects when the environment state changes. The *traffic intensity in state* i is $\rho_i \equiv \lambda_i / s_i \mu_i$. We are interested in cases where $\rho_i > 1$ for at least one i.

The process (X_ϵ, N_ϵ) is an NCD countably-infinite-state CTMC. In particular, (X_ϵ, N_ϵ) has infinitesimal generator matrix A_ϵ which can be expressed in

block form as

$$A_\epsilon = \begin{pmatrix} B_1 + \epsilon Q_{11} I & \epsilon Q_{12} I & \dots & \epsilon Q_{1m} I \\ \epsilon Q_{21} I & B_2 + \epsilon Q_{22} I & \dots & \epsilon Q_{2m} I \\ \epsilon Q_{m1} I & \epsilon Q_{m2} I & \dots & B_m + \epsilon Q_{mm} I \end{pmatrix} \tag{2.2}$$

where I is the (infinite-dimensional) identity matrix and B_i is the infinitesimal generator matrix of the M/M/s_i queue-length process evolving in state i; i.e., with $B_i(j,k)$ being the $(j,k)^{th}$ element of B_i, $B_i(j,j+1) = \lambda_i$ for $j \geq 0$, $B_i(j,j-1) = j\mu_i$ for $1 \leq j \leq s_i$, $B_i(j,j-1) = s_i\mu_i$ for $j \geq s_i$ and diagonal elements chosen so that the row sums are all zero.

The idea is that the CTMC (X_ϵ, N_ϵ) should behave like a Markov-modulated fluid process Y in the time scale $1/\epsilon$ (or, equivalently, when time is sped up by $1/\epsilon$) when space is scaled by ϵ and ϵ is small. The limit process $Y \equiv \{Y(t) : t \geq 0\}$ is specified by the reference CTMC environment process X and deterministic flow rates $r_i \equiv \lambda_i - s_i\mu_i$ for each state i; when $r_i > 0$, there is inflow; when $r_i < 0$, there is outflow. The process X is specified exogeneously and, given a sample path of X, Y evolves according to the differential equation

$$\frac{dY(t)}{dt} = \begin{cases} r_{X(t)} & \text{if} \quad Y(t) > 0 \\ (r_{X(t)})^+ & \text{if} \quad Y(t) = 0 \ , \end{cases} \tag{2.3}$$

where $(x)^+ = \max\{x, 0\}$; e.g., see Kulkarni [24]. The scale-$(1/\epsilon)$ behavior for ϵ small will be formalized by a limit theorem in the next section. It implies that

$$\epsilon N_\epsilon(t/\epsilon) \to Y(t) \text{ w.p.1 as } \epsilon \to 0 \tag{2.4}$$

and leads to the associated approximation

$$N_\epsilon(t) \approx \epsilon^{-1} Y(\epsilon t) \ . \tag{2.5}$$

The statements in (2.4) and (2.5) are stated for a single t, but actually they are for the entire process, i.e., for $t \geq 0$.

The limiting fluid process Y is nontrivial if and only if $\rho_i > 1$ for at least one i; i.e., if $\rho_i \leq 1$ for all i and $Y(0) = 0$, then $Y(t) = 0$ for all t by (2.3). The fluid limit is not difficult to understand: Scaling time by $1/\epsilon$ is tantamount to making the environment processes X_ϵ coincide with the reference environment process X, while accelerating the arrival and service rates. Since the arrival and potential service counting processes are Poisson processes, the limit essentially follows from the strong law of large numbers for the Poisson process.

In our treatment of the queue-length processes N_ϵ each customer's service time is not determined upon arrival but instead by the service rates in effect when that customer is in service. However, in some applications service times may be determined upon arrival. Then it is useful to consider the workload process. Let $W_\epsilon(t)$ represent the workload at time t, i.e., the remaining service time of all customers in the system at time t, for the queue in environment X_ϵ. It turns out that when the workload process $W_\epsilon \equiv \{W_\epsilon(t) : t \geq 0\}$ is properly scaled ($\epsilon W_\epsilon(t/\epsilon)$) it too has a Markov-modulated fluid process limit, just like the queue-length process, but now the deterministic flow rate is $r_i \equiv (\lambda_i/\mu_i) - s_i$ in environment state i.

3. More General Queues in Random Environments

The limiting behavior holds much more generally than for the Markov-modulated M/M/s queue, but some care is needed in formulation. First, the environment process X need not be a CTMC; X can be any finite-state process. It is important to allow for more general reference environment processes,

because the reference environment process strongly determines system behavior. As regularity conditions, we assume that the sample paths of X have only finitely many jumps in any finite interval and are right continuous. We assume that X is specified exogeneously.

It is also not necessary to define the family of slowly changing environment processes X_ϵ directly in terms of the limit X as in (2.1). It suffices to have appropriately normalized versions of X_ϵ converge w.p.1 to X as $\epsilon \downarrow 0$, as indicated below.

We now want to define a queue-length process N in a random environment X. To do so, we need to define N during any time interval $[t, u)$ for which the environment state is i. The idea is that N should act as a general $\mathrm{G/G}/s_i$ queue in state i, where as before s_i is finite and may depend on i, there is unlimited waiting room and the first-come first-served discipline prevails.

To begin the definition of N over $[t, u)$, we need to know the initial conditions at time t. Thus we need to know how many customers are in the system at time t, their residual service times and their order of arrival. These initial conditions can then be treated as a batch arrival at time t. Given the batch arrival at time t determined by the process N before time t, we need to specify the evolution of the general $\mathrm{G/G}/s_i$ model after time t. The generality makes it cumbersome to specify the initial conditions for this process at t. (There would be no difficulty in the Markovian setting of Section 2.) In fact, the limiting behavior should be insensitive to these initial conditions. Hence, we make a general *conditional strong law of large numbers* (CSLLN) assumption on the interarrival times and service times after time t, which should apply to many specific definitions.

We briefly comment on strategies that could be used to make specific def-

initions. One strategy is a single interrupted process for each environment state; i.e., we can have stationary marked point processes of arrival times and service times for each environment state. We can then allocate arrivals from this single process whenever we are in the designated environment state. Thus, we can work with the process $A_i \equiv \{A_i(t) : t \geq 0\}$, where $A_i(t)$ is the number of arrivals in the interval $[0, t]$ where here t is understood to be the total time spent in environment state i. The idea is that we simply continue where we left off when we revisit each environment state.

If we want to assign service times at arrival instants (as with the workload process), then we can start with a stationary *marked* point process of arrival times and service times. Then the service times are the marks that go with the single arrival process above. On the other hand, if we want to let service times be determined while service is in process (as with the queue-length process), then we can allocate service times from a single potential service process associated with each server and each environment state; i.e., allocate service completions from server j in environment state i by the process $S_{ij} \equiv \{S_{ij}(t) : t \geq 0\}$, where $S_{ij}(t)$ is the number of potential service completions in the interval $[0, t]$, where t is the total time server j is busy in environment state i.

A second strategy is an independent restart. Instead of continuing where we left off in a single arrival process when revisiting an environment state, we can start with an independent version of the arrival and service process upon each new visit to an environment state. It is then necessary to specify initial conditions for each restart. For example, we might restart with a stationary version. Note that the Markov-modulated M/M/s queue is consistent with both of these general strategies.

As indicated above, we do not focus on any specific construction. We assume that the queue behavior after each environment process transition is governed by a sequence of *potential* interarrival times and service times, which satisfies a CSLLN, as specified below. However, we do not make any other direct assumptions, such as independence, for this sequence. Let u_{kn} be the potential interarrival time between the $(n-1)^{st}$ and n^{th} new arrival after environment process transition k and let v_{kn} be the n^{th} potential service time. When we work with the queue-length process N, the service times are assigned when customers enter service; when we work with the workload process W, the service times are assigned upon arrival. Let $\mathcal{H}_k$ be the history of the system up to the time of the k^{th} environment process transition. We assume that the environment process is specified exogeneously. Hence, it is reasonable to assume that the future of the environment process is conditionally independent of the potential interarrival-time and service-time sequence $\{(u_{kn}, v_{kn}) : n \geq 1\}$ given the history $\mathcal{H}_k$. We construct the actual interarrival times and service times in the sojourn interval in the new environment state after the k^{th} transition from the potential sequence, using all relevant variables up to transition $k+1$ in the environment process. Given that the k^{th} environment process transition is at time t, we obtain a potential queueing process over the interval $[t, \infty)$. If the next transition of the environment process after time t is at time u, then the resulting queue length process over the interval $[t, u)$ is the restriction of the queueing process on $[t, \infty)$ to the subinterval $[t, u)$. Note that this construction is consistent with both of the two more specific constructions mentioned above.

In this context we assume the CSLLN. For this purpose, let U_{kn} and V_{kn} be the partial sums associated with potential interarrival-time and service-time

sequence $\{(u_{kn}, v_{kn})\}$, i.e.,

$$U_{kn} = u_{k1} + \ldots + u_{kn} \text{ and } V_{kn} = v_{k1} + \ldots + v_{kn} \ . \tag{3.1}$$

We assume that

$$n^{-1}U_{kn} \to \lambda_i^{-1} \text{ and } n^{-1}V_{kn} \to \mu_i^{-1} \text{ w.p.1 as } n \to \infty \tag{3.2}$$

on the set in which the k^{th} environment process transition is to state i, for all k and i. Condition (3.2) implies that the limits are unaffected by conditioning on more of the history $\mathcal{H}_k$.

Note that the CSLLN in (3.2) clearly holds in the case of the M/M/s queue in Section 2 because on the set in which the k^{th} environment process transition is to state i $\{u_{kn} : n \geq 1\}$ and $\{v_{kn} : n \geq 1\}$ are independent sequences of i.i.d. exponential random variables, which are otherwise independent of the history $\mathcal{H}_k$. Similarly, (3.2) holds for the more general GI/GI/s queue because, after some initial segment to represent the initial conditions at the k^{th} environment process transition epoch, on the set in which the k^{th} environment process transition is to a given state, the sequences $\{u_{kn}\}$ and $\{v_{kn}\}$ are again independent sequences of i.i.d. random variables, but now with general distributions. Then the limits in (3.2) hold if and only if these distributions have finite means. The formulation (3.2) allows for even more general models.

To state our limit theorem, let $D[0, \infty)$ be the function space of right-continuous real-valued functions with left limits, endowed with the usual Skorohod topology, as in Ethier and Kurtz [15]. The limit will be convergence w.p.1 for stochastic processes regarded as random elements of $D[0, \infty)$. The proof is given in Section 7.

Theorem 1 *Let $N_\epsilon \equiv \{N_\epsilon(t) : t \geq 0\}$ be the queue-length process and let $W_\epsilon \equiv \{W_\epsilon(t) : t \geq 0\}$ be the workload process of a multi-server queue in*

an exogeneous random environment $X_\epsilon \equiv \{X_\epsilon(t) : t \geq 0\}$. *Assume that the interarrival times and service times in any environment state* i *after any time* t *satisfy the CSLLN in (3.2). Let* $X_\epsilon(\cdot/\epsilon) \to X(\cdot)$ *in* $D[0,\infty)$ *w.p.1 as* $\epsilon \downarrow 0$.

(a) If $\epsilon N_\epsilon(0) \to y$ *in* $\mathbb{R}$ *w.p.1, as* $\epsilon \downarrow 0$, *where* y *is deterministic, and if service times are determined by the environment state when service is performed, then*

$$\epsilon N_\epsilon(\cdot/\epsilon) \to Y(\cdot) \text{ in } D[0,\infty) \text{ w.p.1 as } \epsilon \downarrow 0 \ , \tag{3.3}$$

where Y *is the stochastic fluid process with environment process* X, *deterministic flow rate* $r_i \equiv \lambda_i - s_i\mu_i$ *in state* i *and initial content* $Y(0) = y$.

(b) If $\epsilon W_\epsilon(0) \to z$ *in* $\mathbb{R}$ *w.p.1 as* $\epsilon \downarrow 0$, *where* z *is deterministic, and if service times are assigned upon arrival, then*

$$\epsilon W_\epsilon(\cdot/\epsilon) \to Z(\cdot) \text{ in } D[0,\infty) \text{ w.p.1 as } \epsilon \downarrow 0, \tag{3.4}$$

where Z *is the stochastic fluid limit process with the same environment process* X, *deterministic flow rate* $r_i \equiv (\lambda_i/\mu_i) - s_i$ *in state* i *and initial content* $Z(0) = z$.

4. An MMPP/G/1 Queue Example

In this section we investigate how the stochastic fluid limit in Section 3 behaves as an approximation for the MMPP/G/1 queue when the environment states of the MMPP arrival process change slowly and the queue is unstable in some environment states. We do not consider the Markov-modulated fluid process directly, although it is not difficult to do so. Instead, we consider scaled MMPP/G/1 models and observe when convergence is taking place.

Here we only consider the steady-state workload distribution, which we compute using the algorithm in Choudhury, Lucantoni and Whitt [9], a part

of the Q^2 tool described in Choudhury and Whitt [10]. We could also compute transient distributions using Lucantoni, Choudhury and Whitt [30]. For the special case of exponential service times, the QBD process algorithm of Latouche and Ramaswami [27] is an attractive alternative, but we want to see the effect of the service-time distribution.

Hence, we consider the MMPP/G/1 model with only two environment states, for which the MMPP/G/1 algorithm is very fast. (All computations described below take only a fraction of a second on a SUN SPARC-2 workstation.) It does not seem necessary to consider large models (many environment states) to demonstrate the point that MMPP/G/1 queues can closely approximate stochastic fluid models.

We do the scaling from the outset, so that the environment CTMC can be regarded as fixed. We let the mean holding times in states 1 and 2 be 1 and 5, respectively. The scaling thus appears in the arrival and service rates within the environment states. For each $\epsilon > 0$, we let the mean service time be ϵ, so that the service rate is $1/\epsilon$. We let the arrival rates in environment states 1 and 2 be $2/\epsilon$ and $0.5/\epsilon$, respectively. Hence, the queue is locally unstable in state 1, but locally stable in state 2. The long-run arrival rate is thus $0.75/\epsilon$, so that the traffic intensity is 0.75 for all ϵ. Thus, there exists a proper steady-state distribution.

We consider four different service-time distributions, all having the specified mean: deterministic (D), Erlang of order 4 (E_4), exponential (M) and hyperexponential with balanced means (H_2^b) having SCV $c^2 = 4.0$. (An H_2^b distribution is the mixture of two exponentials, with each exponential contributing equally to the mean. It has two parameters: the mean and the SCV.) Since the deterministic case can cause numerical problems for the numerical

inversion, we use the approximation procedure in Choudhury and Whitt [11] in order to obtain high accuracy. In this setting we could just as well use E_k for large k such as $k = 1024$. (E_{1024} does not increase the algorithm run time noticeably with transform inversion.) Indeed, in this particular example the service-time distribution does not matter much.

Table 1 gives results for the steady-state workload tail probabilities $P(W_\epsilon > x)$ as a function of x, ϵ and the service-time distribution. (Since the scaling was done at the outset, no further scaling is needed.) We consider four values each of x and ϵ: $x = 0.5$, 2.5, 4.5 and 6.5 and $\epsilon = 10^{-k}$ for $k = 1, 2, 3$ and 4.

mean service time ϵ	service-time distribution	tail probability $P(W > x)$			
		$x = 0.5$	$x = 2.5$	$x = 4.5$	$x = 6.5$
10^{-1}	D	0.40260	0.13739	0.04695	0.01604
	E_4	0.41100	0.14350	0.05040	0.01770
	M	0.44246	0.16168	0.06119	0.02316
	$H_2^b,\ c^2 = 4$	0.52216	0.23002	0.01087	0.05168
10^{-2}	D	0.37376	0.11418	0.03488	0.01066
	E_4	0.37383	0.11425	0.03492	0.01067
	M	0.37670	0.11669	0.03614	0.01120
	$H_2^b,\ c^2 = 4$	0.38466	0.12398	0.03997	0.01289
10^{-3}	D	0.37075	0.11183	0.03373	0.01017
	E_4	0.37082	0.11189	0.03376	0.01019
	M	0.37105	0.11208	0.03385	0.01023
	$H_2^b,\ c^2 = 4$	0.37186	0.11281	0.03422	0.01038
10^{-4}	D	0.37044	0.11159	0.03362	0.01013
	E_4	0.37045	0.11160	0.03362	0.01013
	M	0.37047	0.11164	0.03363	0.01013
	$H_2^b, c^2 = 4$	0.37055	0.11169	0.03366	0.01015

Table 1. Steady-state workload tail probabilities $P(W > x)$ in the MMPP/G/1 queue with fixed two-state environment process, mean service time ϵ and over-

all arrival rate $0.75/\epsilon$ as a function of $x,\ \epsilon$ and the service-time distribution.

We also display the first four moments of the steady-state workload W_ϵ as a function of ϵ and the service-time distribution in Table 2. A significant feature of the inversion algorithm [10] is that it can quickly calculate moments of all orders.

The convergence to the fluid limit is evident from Tables 1 and 2, in the way all results rapidly approach common limits as ϵ decreases. For practical

mean service time ϵ	service-time distribution	moments $E[W^k]$			
		$k=1$	$k=2$	$k=3$	$k=4$
10^{-1}	D	1.002	3.65	20.4	151.9
	E_4	1.042	3.88	22.2	170.0
	M	1.161	4.63	28.5	234.6
	$H_2^b,\ c^2=4$	1.621	8.45	67.8	729.0
10^{-2}	D	0.8504	2.860	14.47	97.64
	E_4	0.8508	2.862	14.48	97.76
	M	0.8666	2.941	15.06	102.79
	$H_2^b,\ c^2=4$	0.9134	3.188	16.89	119.36
10^{-3}	D	0.8350	2.786	13.95	93.09
	E_4	0.8354	2.788	13.96	93.21
	M	0.8367	2.794	14.00	93.58
	$H_2^b,\ c^2=4$	0.8414	2.817	14.17	95.04
10^{-4}	D	0.8334	2.779	13.89	92.64
	E_4	0.8335	2.779	13.90	92.65
	M	0.8337	2.779	13.90	92.69
	$H_2^b,\ c^2=4$	0.8341	2.782	13.92	92.83

Table 2. The first four moments of the steady-state workload in the MMPP/G/1 queue with fixed two-state environment process, mean service time ϵ and over-

all arrival rate $0.75/\epsilon$. as a function of ϵ and the service-time distribution.

purposes, the tail probabilities reach their limit by $\epsilon = 10^{-2}$, while the moments reach their limit by $\epsilon = 10^{-3}$. The convergence is recognized by observing when changing ϵ makes negligible difference. From Tables 2 and 3, we see that then the service-time distribution ceases to matter as well, just as we would predict from Theorem 1. (The steady-state distributions of the limiting stochastic fluid model there are independent of the service-time distributions beyond their means.) Theorem 1 does not apply directly, because it describes convergence of sample paths, and thus convergence of finite-dimensional distributions, but not convergence of steady-state distributions. For practical purposes, it is reasonable to interchange the order of limits $\epsilon \downarrow 0$ and $t \to \infty$, even though there is something more to prove.

We do not actually display the distribution for the stochastic fluid limit in Tables 1 and 2, but the numerical results for the queueing models with different scalings clearly show that convergence is taking place as $\epsilon \downarrow 0$ and that the approximation will be excellent for ϵ suitably small. We could use one computed case as an approximation for another.

5. Diffusion Process Refinements

We now obtain refinements to the fluid limits in Theorem 1 in Section 3. For this purpose, we assume a *conditional functional central limit theorem* (CFCLT) for the interarrival times and service times paralleling the CSLLN in (3.2). However, first we need to define what we mean by *conditional weak convergence*. Let $\Rightarrow$ denote the usual weak convergence (convergence in distri-

bution); e.g., see Ethier and Kurtz (1986). Let $\{X_n, n \geq 1\}$ and X be random elements of a separable metric space. Then a standard characterization of weak convergence is: $X_n \Rightarrow X$ as $n \to \infty$ if $Ef(X_n) \to Ef(X)$ as $n \to \infty$ for all bounded continuous real-valued functions f. Now assume that the random elements $X_n, n \geq 1$, and X are also all defined on a common probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. We say that $X_n \Rightarrow X$ conditional on $\mathcal{G}$ if

$$E[f(X_n)|\mathcal{G}] \to E[f(X)|\mathcal{G}] \text{ w.p.1 as } n \to \infty \tag{5.1}$$

for all bounded continuous real-valued functions f. It is elementary that conditional weak convergence implies ordinary weak convergence, because the convergence in (5.1) implies the convergence of the expectations. (Recall that f is bounded and continuous.)

Now we are ready to state our CFCLT assumption. We assume that

$$[n^{-1/2}(U_{k,\lfloor nt \rfloor} - nt\lambda_i^{-1}), n^{-1/2}(V_{k,\lfloor nt \rfloor} - nt\mu_i^{-1})] \Rightarrow [B_1(t), B_2(t)] \tag{5.2}$$

in $D[0,\infty)^2$ as $n \to \infty$, conditional on history $\mathcal{H}_k$, including that the k^{th} environment process transition is to state i, for each i and k, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x and $B_1(t)$ and $B_2(t)$ are two independent zero-drift Brownian motions with diffusion coefficients σ_{ui}^2 and σ_{vi}^2. (It is also straightforward to treat the case in which B_1 and B_2 are dependent.) As with the CSLLN assumption (3.2), it is easy to see that (5.2) is satisfied in the standard special cases.

For simplicity, we assume that we never have $\rho_i = 1$ for any i. Then the diffusion limit applies precisely to the regions that the fluid limit is positive. As in Theorem 1, we assume that the scaled environment processes converge

w.p.1, i.e., $X_\epsilon(\cdot/\epsilon) \to X(\cdot)$ in $D[0,\infty)$ w.p.1 as $\epsilon \downarrow 0$. The diffusion limit depends on both the limiting environment process and the limiting fluid process.

A subtle point is the proper treatment of the times when the fluid process first hits 0. To the right of this time, the diffusion is 0, while to the left of this time, with probability one, it will not be 0. Hence there will be jumps in the limit process at these times. We circumvent this difficulty in the present paper by avoiding these times. We do so by establishing convergence of the finite-dimensional distributions at all times except those times which are jump times with positive probability. This mode of convergence is weaker than weak convergence in $D[0,\infty)$ with one of the usual topologies, but it justifies our desired approximations. We intend to establish a stronger weak convergence result in a subsequent paper.

Let Y be a fluid process and let $\dot{Y}(t-)$ be the left derivative at t, i.e.

$$\dot{Y}(t-) = \lim_{u \downarrow 0} \frac{Y(t) - Y(t-u)}{u} \ , \tag{5.3}$$

which is always well defined since the sample paths of Y are piecewise linear. Then let $\mathcal{T}_Y$ be the set of times t for which

$$P(\dot{Y}(t-) < 0, \ Y(t) = 0) > 0 \ . \tag{5.4}$$

In many applications $\mathcal{T}_Y$ will be empty. For example, if the environment process X is a continuous-time Markov chain, this will be the case.

We avoid these fixed zero-hitting times by considering convergence of the finite-dimensional distributions for all times t not in the designated set $\mathcal{T}_Y$. We say that $Z_\epsilon(\cdot) \Rightarrow_f Z(\cdot)$ in $D[0,\infty)$ as $\epsilon \downarrow 0$ with respect to $\mathcal{T}_Y$ if for all positive integers k and all positive time points $t_1, ..., t_k$ not in $\mathcal{T}_Y$

$$[Z_\epsilon(t_1), ..., Z_\epsilon(t_k)] \Rightarrow [Z(t_1), ..., Z(t_k)] \text{ in } \mathbb{R}^k \text{ as } \epsilon \downarrow 0 \ . \tag{5.5}$$

Obviously convergence in $\Rightarrow_f$ implies ordinary convergence in distribution $Z_\epsilon(t) \Rightarrow Z(t)$ in $\mathbb{R}$ as $\epsilon \downarrow 0$ for a single time point t not in $\mathcal{T}_Y$ (the case $k = 1$ above).

Theorem 2 *Let the processes N_ϵ and W_ϵ be defined as in Theorem 1. Assume that $X_\epsilon(\cdot/\epsilon) \to X(\cdot)$ in $D[0,\infty)$ w.p.1 as $\epsilon \downarrow 0$, $\rho_i \neq 1$ for all i, and the CFCLT in (5.2) holds.*

(a) If $\epsilon N_\epsilon(0) \Rightarrow y$ in $\mathbb{R}$ as $\epsilon \downarrow 0$, where y is deterministic, and if service times are determined by the environment state when service is performed, then

$$\sqrt{\epsilon}(N_\epsilon(\cdot/\epsilon) - Y(\cdot)/\epsilon) \Rightarrow_f \tilde{Y}(\cdot) \text{ in } D[0,\infty) \text{ with respect to } \mathcal{T}_Y \ , \qquad (5.6)$$

where $\tilde{Y}$ is a zero-drift Brownian motion with diffusion coefficient σ_y^2 depending on the state of the limiting fluid and environment process. When $Y(t) = 0$, the diffusion coefficient σ_y^2 of $\tilde{Y}(t)$ is 0; when $Y(t) > 0$ and $X(t) = i$, $\sigma_y^2 = \lambda_i^3 \sigma_{ui}^2 + s_i \mu_i^3 \sigma_{vi}^2$.

(b) If $\epsilon W_\epsilon(0) \Rightarrow z$ in $\mathbb{R}$ as $\epsilon \downarrow 0$, where z is deterministic, and if service times are assigned upon arrival, then

$$\sqrt{\epsilon}(W_\epsilon(\cdot/\epsilon) - Z(\cdot)/\epsilon) \Rightarrow_f \tilde{Z}(\cdot) \text{ in } D[0,\infty) \text{ with respect to } \mathcal{T}_Z \ , \qquad (5.7)$$

where $\tilde{Z}$ is a zero-drift Brownian motion with diffusion coefficient σ_z^2 depending on the state of the limiting fluid and environment process. When $Z(t) = 0$, $\sigma_z^2 = 0$; when $Z(t) > 0$ and $X(t) = i$, $\sigma_z^2 = \lambda_i \sigma_{vi}^2 + \mu_i^{-2} \lambda_i^3 \sigma_{ui}^2$.

6. A Piecewise-Stationary $\mathrm{M_t/G_t/1}$ Queue Example

In this section we investigate how the fluid and diffusion limits in Sections 3 and 5 behave as approximations for the piecewise-stationary $\mathrm{M_t/G_t/1}$ queue.

The piecewise-stationary $M_t/G_t/1$ model is the stationary M/G/1 model in a random environment, where the environment evolves deterministically. We obtain exact results by applying the algorithm to compute the time-dependent workload distribution in Choudhury, Lucantoni and Whitt [8]. Hence we only consider the workload process here.

Let the service-time distribution be fixed; in particular, let it be gamma with mean 1 and squared coefficient of variation (SCV, variance divided by the square of the mean) c^2. (The usual gamma shape parameter is $1/c^2$.) We consider three cases: $c^2 = 0.25,\ 1.0$ and 4.0. There are two environment states. The process is in environment state 1 in the time interval $[0, \epsilon^{-1}]$ and in environment state 2 in the time interval $[\epsilon^{-1}, 2\epsilon^{-1}]$. The arrival rate is 2.0 in environment state 1 and 0.8 in environment state 2. Therefore, the queue is unstable in environment state 1, but stable in environment state 2.

The limit theorems in Sections 3 and 5 imply convergence for the entire workload processes, but for simplicity we consider only the marginal distribution at the end of the second interval. In particular, we look at the distribution of the workload at time $2\epsilon^{-1}$ as a function of ϵ. (Note that ϵ^{-1} is the sojourn time in each environment state.) The fluid process increases at rate 1 in environment state 1 and decreases at rate 0.2 in environment state 2, so that the value of the fluid process is 0.8 at time 2. Theorem 1 implies that

$$\epsilon W_\epsilon(2/\epsilon) \to 0.8 \text{ w.p.1 as } \epsilon \downarrow 0, \tag{6.1}$$

so that the resulting fluid approximation is

$$W_\epsilon(2/\epsilon) \approx 0.8/\epsilon. \tag{6.2}$$

Similarly, Theorem 2 implies that

$$\sqrt{\epsilon}(W_\epsilon(2/\epsilon) - 0.8/\epsilon) \Rightarrow N(0, \sigma^2) \text{ in } \mathbb{R} \text{ as } \epsilon \downarrow 0, \tag{6.3}$$

where $N(m, \sigma)$ is a random variable having the normal distribution with mean m and variance σ^2. (Note that the time 2 is not a zero-hitting time for the limiting fluid process.) By Theorem 2, the diffusion coefficient in environment state i is $\lambda_i(c^2+1)$. By basic properties of Brownian motion, the variance σ^2 in (6.3) is the sum of the variances over the two intervals; i.e., $\sigma^2 = 2.8(c^2+1)$. A consequence of the diffusion approximation is the normal approximation

$$W_\epsilon(2/\epsilon) \approx \frac{0.8}{\epsilon} + \frac{1}{\sqrt{\epsilon}} N(0, 2.8(c^2+1)). \tag{6.4}$$

Equivalently,

$$P\left(W_\epsilon(2/\epsilon) > \frac{0.8}{\epsilon} + x\sqrt{2.8(c^2+1)/\epsilon}\right) \approx P(N(0,1) > x). \tag{6.5}$$

We evaluate the normal approximation (6.5) by computing the exact values of the tail probability on the left side of (6.5) for three values of ϵ, five values of x and three values of c^2, using the numerical inversion algorithm in [8]. The results are shown in Table 3. The numerical inversion computations are accurate to well beyond the given digits, so that the difference between the numerical results and the normal tail probability is the error in the normal approximation.

$1/\epsilon$	c^2	$\sigma/\sqrt{\epsilon}$	$x=-2.0$	$x=-1.0$	$x=0.0$	$x=1.0$	$x=2.0$
10^2	0.25	18.7	0.9817	0.8514	0.5099	0.1671	0.0267
	1.00	23.7	0.9834	0.8530	0.5083	0.1676	0.0281
	4.00	37.4	0.9895	0.8590	0.5067	0.1696	0.0315
10^3	0.25	59.2	0.9787	0.8445	0.5031	0.1613	0.0239
	1.00	74.8	0.9791	0.8447	0.5027	0.1618	0.0245
	4.00	118.3	0.9804	0.8461	0.5023	0.1628	0.0257
10^4	0.25	187.1	0.9775	0.8426	0.5006	0.1600	0.0228
	1.00	236.6	0.9777	0.8424	0.5008	0.1597	0.0233
	4.00	374.2	0.9782	0.8428	0.5007	0.1600	0.0237
$P(N(0,1) > x)$			0.9773	0.8413	0.5000	0.1587	0.0227

Table 3. A comparison of workload tail probabilities for the piecewise-stationary $M_t/G_t/1$ queue with the normal approximation. For each triple (x, ϵ, c^2), the computed workload tail probability is the left side of (6.5).

The results in Table 3 strongly confirm the normal approximation (6.5), which in turn supports the diffusion approximation. The diffusion approximation also confirms the fluid approximation, but in this case of a deterministic environment process the fluid approximation might well be judged as insufficiently accurate. (In Table 3, the fluid approximation yields 1.00 for $x < 0$, 0.50 for $x = 0$ and 0.00 for $x > 0$.) The predicted standard deviation of $W_\epsilon(2/\epsilon)$, $\sigma/\sqrt{\epsilon}$ in (6.3), is displayed in Table 3 to show the spread of the normal distribution approximation about its mean, which is the fluid approximation. For $c^2 = 1$, the predicted standard deviation is 29.6%, 9.4% and 3.0% of the mean $0.8/\epsilon$ when $\epsilon = 10^{-k}$ for $k = 2, 3$ and 4. For applications related to this example, the fluid approximation thus is probably good enough for $\epsilon = 10^{-4}$, but not for $\epsilon = 10^{-2}$.

From Table 3, we see that the quality of the normal approximation improves as ϵ decreases. The quality also tends to improve as c^2 decreases (excluding the case $x = 0$). *Most important, Table 3 shows that the scaling in (6.5) provides a unified view of the different cases considered.*

We remark that the case we consider is a relatively good case for the diffusion approximation. A more difficult case for ϵ not too small would be $\lambda_1 = \rho_1 = 1.1$ and $\lambda_2 = \rho_2 = 0.9$. Then the queue is more likely to become empty (be affected by the barrier at 0) in the first environment state. When the traffic intensities are closer to 1 and the expected environment state sojourn times are not exceptionally long, it is natural to consider RBM with drift, in a

random environment, as an approximating process. That approximation can be obtained as a limit by considering a sequence of models in which the traffic intensities approach 1 in each environment state, as in Iglehart and Whitt [19]. It is considered directly by Asmussen [2] and Karandikar and Kulkarni [20]. We do not consider it here; we intend to compare the different approaches in a future paper.

7. Proofs

In this section we prove Theorems 1 and 2.

Proof of Theorem 1. It should be clear why the processes N_ϵ and W_ϵ are convenient in the two circumstances: If service times are determined by the environment state when service is performed, then the specified service times v_{kn} after environment process transition k will be the service times for the process N_ϵ after time t until the environment state next changes. If, instead, service times are determined upon arrival, then the service times v_{kn} after environment process transition k will be the increments to the workload process W_ϵ until the environment state next changes.

The proofs for parts (a) and (b) are essentially the same, so we primarily focus on part (a). Moreover, the overall argument is similar to previous heavy-traffic limit theorems, e.g., in Iglehart and Whitt [18], [19] and Kella and Whitt [21], so we will be brief. Since the limit process Y has continuous paths, it suffices to use the topology of uniform convergence on bounded intervals on $D[0,\infty)$. We show the desired convergence by focusing on the successive sojourn intervals in environment states.

Let T_n ($T_{\epsilon n}$) be the n^{th} jump time of X (X_ϵ) with $T_n = \infty$ ($T_{\epsilon n} = \infty$) if there are fewer than n jumps. (Usually there will be infinitely many jumps.)

It is not difficult to see that the convergence assumption $X_\epsilon(\cdot/\epsilon) \to X(\cdot)$ in $D[0,\infty)$ w.p.1 as $\epsilon \downarrow 0$ is equivalent to

$$\{(\epsilon T_{\epsilon n}, X_\epsilon(T_{\epsilon n}) : 0 \le n \le n_0\} \to \{(T_n, X(T_n) : 0 \le n \le n_0\} \text{ w.p.1} \qquad (7.1)$$

in $\mathbb{R}^{2n_0+2}$ as $\epsilon \downarrow 0$ for all positive integers n_0. We now condition on the sample path of the limiting environment process and the sample paths indexed by ϵ converging to it.

Note that the fluid process has a constant rate over each sojourn interval, except for those intervals $[T_n, T_{n+1})$ in which the fluid process has a negative rate and hits 0. We break up these intervals into two pieces, so that the fluid process has constant rate on all subintervals. For simplicity, assume that it hits 0 before the endpoint T_{n+1} w.p.1 when it does so, but it is not difficult to treat the other case in which the fluid process hits 0 exactly at T_{n+1} too. We then augment the sequence $\{T_n\}$ by the times that the fluid process first hits the origin in a sojourn interval. Similarly, we augment each sequence $\{\epsilon T_{\epsilon n}\}$ by the times that $\epsilon N_\epsilon(\cdot/\epsilon)$ first hits the origin in a sojourn interval. The fact that the fluid process hits the origin before the end of the interval implies that the scaled queueing process will also for ϵ suitably small. Hence, there is a new time $\epsilon T_{\epsilon n}$ for each new T_n for all sufficiently small ϵ. We can now establish the new version of (7.1) inductively, exploiting the old version of (7.1).

With the new sequence $\{T_n\}$, the fluid process has constant rate on each subinterval $[T_n, T_{n+1})$. Given the w.p.1 representations of (7.1) and the condition $\epsilon N(0) \to y$ w.p.1 as $\epsilon \downarrow 0$, where y is deterministic, it suffices to establish uniform convergence on the environment-state sojourn intervals, i.e., it suffices to show that

$$\sup\{|\epsilon N_\epsilon(t/\epsilon) - \epsilon N_\epsilon(T_{\epsilon n}) - \gamma_i t| : \epsilon T_{\epsilon,n} \le t < \epsilon T_{\epsilon,n+1}\} \to 0 \text{ w.p.1 as } \epsilon \downarrow 0 \quad (7.2)$$

on the set $\{X(T_n) = i\}$ for each i and n, where $\gamma_i = 0$ if $Y(t) = 0$, $T_n \leq t \leq T_{n+1}$ (which will occur if $Y(T_n) = 0$, $X(T_n) = i$ and $r_i \leq 0$), and $\gamma_i = r_i$ otherwise. Note that, for almost all sample points in the set $\{X(T_n) = i\}$, $X_\epsilon(\epsilon T_{\epsilon n}) = i$ as well for all ϵ suitably small. On the set $\{X(T_n) = i\}$ and over the interval $T_{\epsilon n} \leq t < T_{\epsilon,n+1}$, (7.2) corresponds to a heavy-traffic functional strong law of large numbers (FSLLN) for the queue-length process N_ϵ. This heavy-traffic FSLLN follows by the same argument as for the heavy-traffic functional central limit theorem (FCLT) in Iglehart and Whitt [18], assuming that the interarrival times and service times satisfy FSLLNs. (That heavy-traffic argument in [18] has some technical complications: In [18] it is shown that it suffices to assume that all servers are continuously busy when the traffic intensity is greater than or equal to 1. See Chen and Mandelbaum [4] for related arguments applied to networks of single-server queues.) We also need a functional generalization of (3.2). However, the assumed ordinary CSLLNs in (3.2) are actually equivalent to the associated conditional FSLLNs; see Theorem 4 of Glynn and Whitt [16]. Hence (3.3) is proved. When we turn to establishing (3.4), we use the fact that the CSLLN assumption (3.2) implies a corresponding CSLLN for the total input $\sum_{i=1}^{A(t)} v_i$, where $A(t)$ is the arrival counting process associated with the interarrival-time sequence $\{u_n\}$.

Proof of Theorem 2. The proof begins just as in the proof of Theorem 1. Once we specify the limiting environment process sample path, we specify the corresponding fluid process sample path. Then we can recursively apply the heavy-traffic limit theorem in Iglehart and Whitt [18] over successive environment process sojourn intervals. Suppose that, after the conditioning, $[T_i, T_{i+1})$ is an interval for which $Y(T_{i+1}-) > 0$ and $Y(T_{i+1}) = 0$. The previous heavy-traffic argument establishes a FCLT over the intervals $[T_i, t]$ for any t such that

$T_i < t < T_{i+1}$. This in turn is equivalent to weak convergence in the space $D([T_i, T_{i+1}))$, i.e., where the interval $[T_i, T_{i+1})$ is open on the right; see [42]. This means that we do not establish convergence at T_{i+1}, but we do not need to because w.p.1 $T_{i+1} \in \mathcal{T}_{Y_j}$. (We discuss this point further below.) Even though we do not establish convergence at zero-hitting times, we can proceed inductively over successive intervals $[T_i, T_{i+1}]$. The initial position for the diffusion is determined by the previous interval, with the future increments of the Brownian motion independent of the history in previous intervals, by virtue of (5.2). The diffusion coefficients are obtained from p. 155 of Iglehart and Whitt [18] and Section 2 of Whitt [40]. The argument just given establishes conditional weak convergence as in (5.1) for the finite dimensional distributions, excluding time points in $\mathcal{T}_Y$, where the conditioning σ-field $\mathcal{G}$ in (5.1) represents the limiting environment process. As indicated after (5.1), this conditional weak convergence of the finite-dimensional distributions directly implies ordinary weak convergence of the finite-dimensional distributions, which is the result to be proved.

References

[1] 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.

[2] Asmussen, S. (1995) Stationary distribution for fluid flow models with or without Brownian noise. *Stochastic Models*, 11, 21–49.

[3] Chang, C.-S. and Nelson, R. (1993) Perturbation analysis of the M/M/1 queue in a Markovian environment via the matrix-geometric method.

Stochastic Models 9, 233–246.

[4] Chen, H. and Mandelbaum, A. (1991a) Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Opns. Res.* 16, 408–446.

[5] Chen, H. and Mandelbaum, A. (1991b) Discrete flow networks, diffusion approximations and bottlenecks. *Ann. Probab.* 19, 1463–1519.

[6] Chen, H. and Whitt, W. (1993) Diffusion approximations for open queueing networks with service interruptions. *Queueing Systems* 13, 335–359.

[7] Chen, H. and Yao, D. D. (1992) A fluid model for systems with random disruptions. *Opns. Res.* 40, S239–S247.

[8] Choudhury, G. L. Lucantoni, D. M. and Whitt, W. (1993) Numerical solution of piecewise-stationary $M_t/G_t/1$ queues. *Opns. Res.*, to appear.

[9] Choudhury, G. L., Lucantoni and Whitt, W. (1995) Squeezing the most out of ATM. *IEEE. Trans. Commun.* 44, 203–217.

[10] Choudhury, G. L. and Whitt, W. (1995a) Q^2: A new performance analysis tool exploiting numerical transform inversion. *Proc. Third Int. Workshop on Modelling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS '95*, Durham, NC, January, 1995) 411–415.

[11] Choudhury, G. L. and Whitt, W. (1995b) Non-probability approximations for probability distributions to aid numerical transform inversion. in preparation.

[12] Courtois, P. (1977) *Decomposability*, Academic Press, New York.

[13] de Smit, J. H. A. (1986) The single server semi-Markov queue. *Stoch. Proc. Appl.* 22, 37–50.

[14] Elwalid, A. I. and Mitra, D. (1991) Analysis and design of rate based congestion control of high speed networks, I: stochastic fluid models, access regulation. *Queueing Systems* 9, 29–64.

[15] Ethier, S. and Kurtz, T. (1986) *Markov Processes: Characterization and Convergence*, Wiley, New York.

[16] Glynn, P. W. and Whitt, W. (1988) Ordinary CLT and WLLN versions of $L = \lambda W$. *Math. Opns. Res.* 13, 674–692.

[17] Iglehart, D. L. (1971) Multiple channel queues in heavy traffic, IV: law of the iterated logarithm. *Z. Wahrscheinlichkeitsth. verw. Geb.* 17, 168–180.

[18] Iglehart, D. L. and Whitt, W. (1970a) Multiple channel queues in heavy traffic, I *Adv. in Appl. Prob.* 2, 150–177.

[19] Iglehart, D. L. and Whitt, W. (1970b) Multiple channel queues in heavy traffic, II: sequences, network and batches. *Adv. Appl. Prob.* 2, 355–369.

[20] Karandikar, R. L. and Kulkarni, V. G. (1995) Second-order fluid flow models: reflected Brownian motion in a random environment. *Opns. Res.*, 45, 77–88.

[21] Kella, O. and Whitt, W. (1990) Diffusion approximations for queues with server vacations. *Adv. Appl. Prob.* 22, 706–729.

[22] Kella, O. and Whitt, W. (1992a). A storage model with a two-state random environment. *Opns. Res.* 40, S257–S262.

[23] Kella, O. and Whitt, W. (1992b). A tandem fluid network with Lévy input, in *Queueing and Related Models*, U. N. Bhat and I. V. Basawa (eds.), Oxford Science Publications, Clarendon Press, Oxford, 112–128.

[24] Kulkarni, V. G. (1995) Fluid models for single buffer systems. in *Frontiers of Queueing Theory*, J. Dshalalow (ed.), to appear.

[25] Kushner, H. J. and Dupuis, P. G. (1992) *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York.

[26] Latouche, G. (1991) First passage times in nearly decomposable Markov chains. *Numerical Solutions for Markov Chains*, W. J. Stewart (ed.), Marcel Dekker, New York, 401–411.

[27] Latouche, G. and Ramaswami, V. (1993) A logarithmic reduction algorithm for quasi-birth-death processes. *J. Appl. Prob.* 30, 650–674.

[28] Latouche and Schweitzer (1995) A Markov modulated nearly completely decomposable M/M/1 queue. *Proceedings of the Second International Workshop on the Numerical Solution of Markov Chains*, W. J. Stewart (ed.), to appear.

[29] Lucantoni, D. M. (1993) The BMAP/G/1 queue: a tutorial. *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, L. Donatiello and R. Nelson (eds.), Springer, New York, 330–358.

[30] Lucantoni, D. M., Choudhury, G. L. and Whitt, W. (1994) The transient BMAP/G/1 queue. *Stochastic Models* 10, 145–182.

[31] Mandelbaum, A. and Massey, W. A. (1995) Strong approximations for time-dependent queues. *Math. Opns. Res.* 20, 33–64.

[32] Mitra, D. (1988) Stochastic theory of a fluid model of multiple failure-susceptible producers and consumers coupled by a buffer. *Adv. Appl. Prob.* 20, 646–676.

[33] Neuts, M. F. (1981) *Matrix-Geometric Solutions in Stochastic Models*, The Johns Hopkins University Press, Baltimore.

[34] Philippe, B., Saad, Y. and Stewart, W. J. (1992) Numerical methods in Markov chains. *Opns. Res.* 40, 1156–1179.

[35] Regterschot, G. J. K. and de Smit, J. H. A. (1986a) The queue M/G/1 with Markov modulated arrivals and services. *Math. Opns. Res.* 11, 465–483.

[36] Regterschot, G. J. K. and de Smit, J. H. A. (1986b) A semi-Markov queue with exponential service times. *Semi-Markov Models Theory and Applications* J. Janssen (ed.) Plenum Press, New York, 369–382.

[37] Rogers, L. C. G. (1994) Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *Ann. Appl. Prob.* 4, 390–413.

[38] Simon, H. A. and Ando, A. (1961) Aggregation of variables in dynamic systems. *Econometrica* 29, 111–138.

[39] Stern, T. E. and Elwalid, A. I. (1991) Analysis of separable Markov-modulated rate models for information handling systems. *Adv. Appl. Prob.* 23, 105–139.

[40] Whitt, W. (1971) Weak convergence theorems for priority queues: preemptive-resume discipline. *J. Appl. Prob.* 8, 74–94.

[41] Whitt, W. (1973) Heavy traffic limit theorems for queues: a survey. In *Mathematical Methods in Queueing Theory*, A. B. Clarke (ed.), Springer-Verlag, New York, 307–350.

[42] Whitt, W. (1980) Some useful functions for functional limit theorems. *Math. Opns. Res.* 5, 67–85.