Discrete Event Dynamic Systems: Theory and Applications 2, (1992): 7-60

The Brownian Approximation for Rate-Control Throttles and the G/G/1/C Queue

ARTHUR W. BERGER
AT&T Bell Laboratories, Room 3H-601, Holmdel, NJ 07733-3030

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

Received June 13, 1991; Revised December 23, 1991

Abstract. This paper studies approximations to describe the performance of a rate-control throttle based on a token bank, which is closely related to the standard G/G/1/C queue and the two-node cyclic network of ·/G/1/∞ queues. Several different approximations for the throttle are considered, but most attention is given to a Brownian or diffusion approximation. The Brownian approximation is supported by a heavy-traffic limit theorem (as the traffic intensity approaches the upper limit for stability) for which an upper bound on the rate of convergence is established. Means and squared coefficients of variation associated with renewal-process approximations for the overflow processes are also obtained from the Brownian approximation. The accuracy of the Brownian approximation is investigated by making numerical comparisons with exact values. The relatively simple Brownian approximation for the job overflow rate is not very accurate for small overflow rates, but it nevertheless provides important insights into the way the throttle design parameters should depend on the arrival-process characteristics in order to achieve a specified overflow rate. This simple approximation also provides estimates of the sensitivity of the overflow rates to the model parameters.

Key Words: rate-control throttle, token bank, leaky bucket, communication networks, broadband integrated services digital networks (B-ISDNs), asynchronous transfer mode (ATM), G/G/1/C queues, queues with finite waiting rooms, overflow processes, Brownian models, diffusion approximations, heavy traffic, parametric-decomposition approximations

1. Introduction

In this paper we have two principal objectives and thus two potential audiences. Our first objective is to gain a better understanding of the performance of rate-control throttles, which regulate the admission of jobs (or customers) to some (unspecified) system. It will be clear that a rate-control throttle is a discrete-event dynamic system (DEDS). Our second objective is to gain a better understanding of Brownian or diffusion approximations for queueing models. Readers with interest in only one of these topics may thus want to skip some material.

We were motivated to conduct this study because we were studying the performance of a multiclass rate-control throttle based on token banks, which was suggested by our colleague R. Milito, and indeed the results here are applied to analyze the multiclass throttle in Berger and Whitt [1990, 1992]. In the multiclass throttle, the admission of jobs is regulated by token banks dedicated to each class and a single, shared overflow token bank.

To make our motivation clear, we first describe the multiclass throttle in Berger and Whitt [1990,1992] in more detail. There are N classes indexed by i. The token bank dedicated to class i is bank i, and the overflow token bank is bank 0. On the arrival of a job of class i, if bank i contains a token, then the job is admitted and a token is removed from bank i; if bank i is empty, then the job overflows to bank 0, where it gets a second chance to be admitted. If bank 0 contains a token, then the job is admitted and a token is removed from bank 0. If both banks i and 0 are empty, then the job is rejected by the throttle. (Herein, we consider the job to be lost, though alternatively it could be queued, or marked and admitted, and then treated as a lower-priority class.) Class i tokens arrive deterministically, evenly spaced, at a rate r_i, to bank i of finite capacity C_i. (In practice, token banks are typically implemented as counters.) If bank i is full, then the token overflows to bank 0 of capacity C_0. If both banks i and 0 are full, then the token is dropped and lost.

The decision maker chooses the parameters $r_1, \ldots, r_N$ and $C_0, C_1, \ldots, C_N$. The rates r_i determine the maximum, sustained, admission rate that is guaranteed for each class. The maximum sustained admission rate for all classes is $\Sigma_{i=1}^{N} r_i$. The overflow bank allows the excess capacity for some classes to be used by other classes. If class-i jobs arrive deterministically at rate λ_i, evenly spaced like tokens, then the admission rate for class- i jobs without excess capacity is indeed $\min\{r_i, \lambda_i\}$. However, class-i jobs will typically arrive randomly, so that the admission rate for class-i jobs will typically be somewhat less than $\min\{r_i, \lambda_i\}$ when there is no excess capacity. The capacities C_i limit the instantaneous burst of arrivals that may be admitted. A larger capacity C_i makes the long-run average admission rate for class-i jobs without excess capacity closer to $\min\{r_i, \lambda_i\}$, but allows for larger bursts of arrivals in the short run. Assuming that the jobs arrive according to stochastic processes, both the token rates and the bank capacities influence the steady-state-per-class blocking and throughput, as well as the transient response.

In this paper we focus on the performance of a single token bank. This bank could be either a dedicated bank (in a single-class or multiclass throttle) or an overflow bank. In particular, our model of a token bank in this paper has jobs and tokens arriving in two separate streams. The token arrival process is typically deterministic, but we allow more general token arrival processes, as would occur at an overflow bank. We assume that the job and token arrival processes are stochastically independent (which for the case of an overflow bank would be a simplifying assumption). The tokens are put in a bank of capacity C, overflowing when the bank is full. Arriving jobs finding a token in the bank are admitted, with each taking a token along. (Thus token admissions coincide with job admissions.) When there are no tokens in the bank upon a job arrival, the job overflows. We are interested in blocking probabilities, the admission processes and the overflow processes for jobs and tokens.

Since we are interested in the multiclass throttle, we are especially interested in the job and token overflow streams from the dedicated banks. In particular, our primary goal is to develop approximations for the job and token overflow processes. We determine approximations for the mean, the squared coefficient of variation (SCV, variance divided by the square of the mean), and even the full distribution of the time between overflows associated with renewal-process approximations for the overflow streams. These renewal-process approximations are used in Berger and Whitt [1990, 1992] to analyze the multiclass throttle using a parametric-decomposition approximation.

In this paper, we consider several different approximations for the token bank, but we primarily investigate the Brownian or diffusion approximation, as contained in Chapter 5 of Harrison [1985] and Chapters 7 and 8 of Newell [1982]. Note that the Brownian approximation for the token bank is a continuous approximation for a DEDS, which has appeal because relatively classical methods can be applied. Other work on Brownian or diffusion approximations for finite-capacity single-server queues is contained in Section 6.8 of Whitt [1969], Sweet and Hardin [1970], Kennedy [1973], Gaver and Shedler [1973a,b], Gelenbe [1975], Kimura, Ohno, and Mine [1979], Gelenbe and Mitrani [1980], Kimura [1985], Yao and Buzacott [1985a,b], Coffmann, Puhalsky, and Reiman [1991], Fendick and Rodrigues [1991], and Dai and Harrison [1991]. Via the connection to the two-node cyclic (CQN), this Brownian approximation is also related to Chen and Mandelbaum [1991a,b].

We contribute to a better understanding of the Brownian approximation in several ways: First, we derive the asymptotic variance constants for the Brownian barrier-regulator or local-time processes, which provide SCVs for the interoverflow times in renewal-process approximations for the overflow processes (Theorem 4.1). Second, we establish an upper bound on the rate of convergence in the supporting heavy-traffic limit theorem (Theorem 4.4). Third, we investigate the accuracy of the approximations by making numerical comparisons (Section 8). Finally, we show how the relatively simple Brownian approximation for the job overflow rate can provide important insights into the way system design parameters should depend on the arrival process characteristics in order to achieve specified overflow rates (Section 9). It also permits us to do a rough sensitivity analysis, e.g., to see how the accuracy of estimates of the overflow rates depends on the accuracy of estimates of the arrival process variability. Readers primarily interested in rate-control throttles may want to skip the limit theorems in Section 4.5.

In developing the multiclass throttle described above, we were primarily motivated by the desire to manage call setup requests in telecommunication switching systems, where different classes may be different types of calls such as line or trunk originations. Then the job arrival processes might be well modeled by Poisson processes, and interest centers on designs to achieve blocking probabilities of order 10^{-3} to 10^{-1}. For these applications, single-bank rate-control throttles were previously investigated by Doshi and Heffes [1983], Eisenberg [1983], and Berger [1991a,b].

Rate-control throttles also have potential for high-speed communication networks, such as broadband integrated services digital networks (B-ISDNs), where the throttle regulates the admission of asynchronous transfer mode (ATM) cells. Then the job arrival process may differ substantially from Poisson processes and interest may center on job- (cell-) blocking probabilities of order 10^{-6} to 10^{-9}. Turner [1986] suggested a throttle based on the leaky bucket to monitor the admission of ATM cells. Motivated by the interest in BISDNs, many researchers have subsequently examined (single-class) rate-control throttles based on token banks or leaky buckets; e.g., see Eckberg, Luan, and Lucantoni [1989], Sidi et al. [1989], Sohraby and Sidi [1990], Rathgeb [1990], Kroner, Theimer, and Briem [1990], Budka and Yao [1990], Budka [1990], and Elwalid and Mitra [1991]. The multiclass rate-control throttle investigated in Berger and Whitt [1990, 1991] is also applicable to BISDN/ATM, with the different classes representing different virtual channels that share a common virtual path.

Our analysis here indicates that the Brownian approximation, or any other approximation based on a single-parameter characterization of variability, is unlikely to be able to accurately predict such small blocking probabilities as 10^{-6} to 10^{-9} when heavy-traffic conditions do not prevail. Indeed, our results (exact numerical results, not approximations or simulations) clearly demonstrate this. We show that small blocking probabilities can be dramatically different if a renewal job arrival process is replaced by another with the same first two interarrival-time moments; e.g., in the last row of Table 6 one exact blocking probability is 0.11×10^{-4}, while the other exact blocking probability is 0.72×10^{-15}.

Nevertheless we contend that the relatively simple Brownian approximation for the overflow rate (in (25)) can provide important insights into system design and performance. In particular, from this simple formula we can see, at least roughly, how the design parameters should depend on the arrival-process characteristics to achieve specified blocking probabilities. We can also perform sensitivity analyses; see Section 9.

A single dedicated token bank is very closely related to other models, including the leaky-bucket rate-control mechanism. In Section 7 we define the leaky bucket precisely and show that it is equivalent to a modified dedicated bank in which the deterministic token arrival stream is turned off whenever the bank becomes full of tokens, and turned on again upon the next job arrival. We also compare the leaky bucket to the token bank in Section 7. (Related work appears in Budka [1990].) The difference between the token-bank and leaky-bucket throttles can be significant when the token bank has very small capacity (e.g., 1), but for larger capacities the performance of the two systems tends to be essentially the same. Thus, we regard our performance analysis as being applicable to either rate-control scheme.

The single bank is also closely related to the G/G/1/C queue and the two-node cyclic network of $\cdot$/G/1/∞ queues with C customers. The idea is to think of the G/G/1/C as representing the token bank, so that the arrival process in the queueing model is the token arrival process, and the service process in the queueing model is constructed from the job arrival process. However, note that the server works in the G/G/1/C model only when customers are present, whereas the job arrival process runs continuously in the token bank model. The token bank would be equivalent to a G/G/1/C queue with deterministic arrival process if we identified a job arrival in the token bank with a *potential* service completion in the G/G/1/C queue, or if we turned off the job arrival process in the token bank model whenever the bank becomes empty and turned it on again upon a token arrival. (The token bank thus directly corresponds to a queue with autonomous service as discussed in Chapter 8 of Borovkov [1976].)

Similarly, the token bank would be equivalent to a two-node cyclic network of $\cdot$/G/1/∞ queues, one of which is $\cdot$/D/1/∞, if in the token bank model we turned off the token arrival process when the bank is full as well as the job arrival process when the bank is empty. (Here and above "turn off" means suspending time for the given arrival process.) In the M/M case (two independent Poisson streams), the token bank, the leaky bucket, the G/G/1/C queue, and the two-node cyclic CQN are all stochastically equivalent because of the lack of memory property associated with the exponential distribution. The Brownian approximation obtained via a heavy-traffic limit theorem is the same for all four models, too. (Thus, differences in the values of corresponding performance measures for these models indicate limitations in the numerical accuracy of the Brownian approximations.)

In this paper, we consider four different approximations for the token-bank rate-control throttle, with each successive approximation involving more detail. In Section 2 we introduce a simple "first-order" *fluid approximation*; it is a deterministic approximation based only on the job and token arrival rates. We find that it is remarkably accurate for describing the throughputs and overflow rates when the job and token arrival rates are not nearly equal. In Section 3 we introduce a "second-order" *Poisson approximation*; it is the M/M/1/C model, representing the stochastic nature of the arrival processes as well as the rates. We derive the SCV of the overflow process for an M/M/1/C queue (Theorem 3.1) and use it to determine the asymptotic variance constants of the Brownian boundary regulator processes, which in turn provide the SCVs partially characterizing renewal-process approximations for the overflow processes.

In Section 4 we introduce the "third-order" *Brownian* or *diffusion approximation*; it exploits the degree of variability (the SCVs) of each arrival stream, as well as the stochastic nature. In Section 5 we discuss a "fourth-order" *Markov chain approximation*; it is an exact Markov chain analysis of the GI/M^X/1/C model, having a renewal token arrival process and a batch-Poisson job arrival process, or the GI/PH/1/C model, having a phase-type service time distribution. For models which are not one of these models, we fit the token arrival stream to a convenient renewal (GI) stream and the job arrival stream to a batch Poisson (M^X) stream or phase-type (PH) renewal stream by matching moments. For a dedicated bank, the token arrival stream would be deterministic (D). Otherwise, we suggest fitting a simple phase-type renewal arrival process.

In Section 6 we further discuss approximations for the SCVs of the job and token overflow streams. We do an asymptotic analysis for non-heavy-traffic regimes that provides additional support for the Brownian approximation for the overflow SCVs. We also investigate the properties of the accepted job stream. In Section 7 we compare the token bank to the leaky bucket. In Section 8 we evaluate the performance of the approximations by making numerical comparisons. In Section 8.5 we introduce a new refinement to the Brownian approximation, which is especially effective for the M/M/1/C model. In particular, motivated by continuous-distribution approximations for discrete distributions, we consider the Brownian model with barriers at $-1/2$ and $C + 1/2$ instead of at 0 and C. In Section 9 we discuss insights that can be gained from the relatively simple Brownian approximation for the job overflow rate. For example, we do sensitivity analysis. Finally, in Section 10 we state our conclusions.

2. The Fluid Approximation

The first approximation scheme is a simple deterministic fluid model: We act as if both jobs and tokens arrive at the dedicated bank deterministically and continuously like a fluid at constant rates, i.e., we assume that the jobs and tokens arrive at rates λ and r, respectively, and let the overflow rates be

$$\lambda' = \lambda - \min\{\lambda, r\} \quad \text{and} \quad r' = r - \min\{\lambda, r\}. \tag{1}$$

This fluid approximation yields no useful second parameters for the overflow streams.

The fluid approximation in (1) is much more elementary than some other finite-buffer fluid models, e.g., in Anick, Mitra, and Sondhi [1982] and Elwalid and Mitra [1991]. In these other fluid models, the arrival rates are governed by stochastic processes; i.e., they involve fluid flow in a random environment. Here we consider these fluid approximations for the relatively trivial special case in which the environment is fixed. The numerical results for this simple case provide some additional insight into the performance of the more elaborate fluid approximations.

There is a useful *conservation law* for the token bank, which we will exploit later. Since admitted jobs must be matched with admitted tokens, the long-run job admission rate must equal the long-run token admission rate; i.e., in any token bank we must have

$$\lambda - \lambda' = r - r'. \tag{2}$$

Note that (1) satisfies (2).

3. The Poisson Approximation

The second scheme is a Poisson approximation: We act as if both the token and job arrival streams are Poisson processes with the given rates. Then the token bank is stochastically equivalent to the classical M/M/1/C queue with 1 server and $C - 1$ extra waiting spaces. To relate the job arrival process in the token bank model to the service process in the M/M/1/C queue, we regard the job arrivals as potential service completions in the M/M/1/C queue. If a customer (token) is present in the M/M/1/C queue, then the job arrival causes a real service completion and a departure in the M/M/1/C model. However, if there is no customer present upon this job arrival, then this job arrival has no effect. Since the job arrival process is a Poisson process, the distribution of the time until the first service completion after the next customer (token) arrival is still exponential as it should be.

Let $p(n)$ be the probability that the number of tokens in the bank is n at an arbitrary time in equilibrium, and let $\rho = r/\lambda$. Then, from standard M/M/1/C formulas,

$$p(n) = \begin{cases} \dfrac{1}{C+1} & \text{if } \rho = 1, \\ \dfrac{(1-\rho)\rho^n}{1-\rho^{C+1}} & \text{if } \rho \neq 1 \end{cases} \tag{3}$$

for $0 \leq n \leq C$. The associated mean is

$$\sum_{n=1}^{C} np(n) = \begin{cases} \dfrac{C}{2} & \text{if } \rho = 1, \\ \dfrac{\rho}{1-\rho} - \dfrac{(C+1)\rho^{C+1}}{1-\rho^{C+1}} & \text{if } \rho \neq 1. \end{cases} \tag{4}$$

Since Poisson arrivals see time averages (PASTA), see Wolff [1982], the job-blocking probability coincides with the probability of emptiness in an M/M/1/C model with traffic intensity $\rho = r/\lambda$. Likewise, the probability that a token is blocked equals the probability an M/M/1/C system with traffic intensity ρ is full or, equivalently, the probability that an M/M/1/C system with traffic intensity ρ^{-1} is empty. Hence, we have the following explicit expressions for the overflow probabilities

$$\frac{\lambda'}{\lambda} = \begin{cases} \dfrac{1}{C+1} & \text{if } \rho = 1, \\ \dfrac{1-\rho}{1-\rho^{C+1}} & \text{if } \rho \neq 1, \end{cases} \tag{5}$$

and

$$\frac{r'}{r} = \begin{cases} \dfrac{1}{C+1} & \text{if } \rho = 1, \\ \dfrac{1-\rho^{-1}}{1-\rho^{-(C+1)}} & \text{if } \rho \neq 1. \end{cases}$$

From (5), it is easy to check that the conservation law (2) is satisfied.

Since the overflow processes are renewal processes under the Poisson assumption, it is natural to use the SCV of an inter-renewal interval as a second parameter to partially characterize the overflow streams. Let c_J^2 and c_T^2 be the SCVs of the job and token overflow processes, respectively. We obtain closed-form expressions for these SCVs, which seem to be new. It is remarkable that $c_J^2 = c_T^2$. Unfortunately, we do not yet have a good intuitive explanation.

THEOREM 3.1. For the M/M/1/C model,

$$c_J^2 = c_T^2 = \begin{cases} \dfrac{2C^2+4C+3}{3C+3} & \text{if } \rho = 1, \\ \dfrac{(1+\rho)(1-\rho^{2C+2}) - 4(C+1)(1-\rho)\rho^{C+1}}{(1-\rho)(1-\rho^{C+1})^2} & \text{if } \rho \neq 1. \end{cases} \tag{6}$$

Proof. By direct calculation, we derive (6) for c_J^2. Then c_T^2 is also given by (6) but with ρ^{-1} substituted for ρ. However, (6) remains unchanged when ρ is replaced by ρ^{-1}. To derive (6) and higher moments, let J be the time between successive job overflows and let B be the length of a busy period for tokens (the interval beginning when a token arrives at an empty token bank until the bank is next empty again). Then

$$J \stackrel{d}{=} X + (1 - I)(B + J), \tag{7}$$

where $\stackrel{d}{=}$ denotes equality in distribution, the four random variables on the right are independent, X is exponential with mean $1/(\lambda + r)$, and $P(I = 1) = 1 - P(I = 0) = \lambda/(\lambda + r)$. We obtain (7) by considering what happens after a job overflow. Because of the continuous-time Markov chain structure, the time until the next event is the exponential random variable X. Then $I = 1$ if the next event is a job arrival, in which case $J = X$. If $I = 0$, then the next event is a token arrival. Then the remaining time until the next job overflow is the sum of a token busy period B plus the time until the next job overflow after the system becomes empty (which is distributed the same as J). The Markov property implies that B and J are independent on the right in (7). (It is easy to see that $J \stackrel{d}{=} B$ when $C = \infty$, because then B satisfies the same relation (7), with the two B variables on the right being independent.)

From (7), we obtain

$$\lambda E[J] = 1 + rE[B] \tag{8}$$

and

$$\lambda^2 E[J^2] = r\lambda E[B^2] + 2r^2(E[B])^2 + 4rE[B] + 2. \tag{9}$$

To find $E[B]$ and $E[B^2]$, we uniformize the continuous-time Markov chain, see Keilson [1979], and apply discrete-time Markov chain (MC) formulas from Kemeny and Snell [1959]. In particular, we use representation

$$B \stackrel{d}{=} \sum_{j=1}^{D} X_j, \tag{10}$$

where X_j, $j \geq 1$, are i.i.d. exponential random variables with mean $1/(\lambda + r)$ and D is the number of steps until absorption into state 0 from state 1 in an associated discrete-time MC. The absorbing discrete-time MC is a simple random walk on $\{1, 2, \ldots, C\}$ which takes a step up with probability $r/(\lambda+r)$ and a step down with probability $\lambda/(\lambda + r)$. In state C, instead of going up, the walk stays in state C with probability $r/(\lambda + r)$; in state 1 the walk is absorbed, instead of going down, with probability $\lambda/(\lambda + r)$.

From (10), we obtain

$$E[B] = \frac{E[D]}{\lambda + r} \quad \text{and} \quad E[B^2] = \frac{E[D^2] + E[D]}{(\lambda + r)^2}. \tag{11}$$

Then

$$E[D] = \sum_{j=1}^{C} N_{1j} \quad \text{and} \quad E[D^2] = \sum_{j=1}^{C} (2N^2 - N)_{1j} \tag{12}$$

where $N = (I - Q)^{-1}$ is the fundamental matrix associated with the absorbing MC on $\{1, \ldots, C\}$ with transition matrix Q, see p. 49 of Kemeny and Snell [1959]. We obtain

(6) by exploiting the special random walk structure, see p. 149 of Kemeny and Snell [1959]. In particular, we obtain

$$E[D] = \begin{cases} 2C & \text{if } \rho = 1, \\ \dfrac{1-\rho^C}{(1-\rho)\lambda} & \text{if } \rho \neq 1, \end{cases} \tag{13}$$

$$E[D^2] = \begin{cases} \dfrac{(4C+1)(2C+1)}{3(\lambda+r)^2} & \text{if } \rho = 1, \\ \dfrac{2(1+\rho)^2}{(1-\rho)^3}\left(1-(2C+1)\rho^C+(2C+1)\rho^{C+1}-\rho^{2C+1}\right) & \text{if } \rho \neq 1, \end{cases} \tag{14}$$

and

$$E[B^2] = \frac{2}{\lambda^2(1-\rho)^3}\left(1-(2C+1)\rho^C(1-\rho)-\rho^{2C+1}\right), \tag{15}$$

from which (6) follows. ■

Thus, for the M/M/1/C model we can nicely characterize the SCVs of the overflow streams. From Theorem 3.1 we can deduce the following properties. We write $g(x) \sim f(x)$ as $x \to \infty$ if $g(x)/f(x) \to 1$ as $x \to \infty$.

COROLLARY. For the M/M/1/C model, c_f^2 is increasing in C, increasing in ρ for $\rho \leq 1$ and decreasing in ρ for $\rho \geq 1$. As $C \to \infty$, $c_f^2 \to (1+\rho)/|1-\rho|$ for $\rho \neq 1$. As $\rho \to 0$ and as $\rho \to \infty$, $c_f^2 \to 1$.

4. The Brownian Approximation

The third scheme incorporates the SCVs of the token and job arrival processes in the analysis. Since the token arrival process at the dedicated bank is deterministic, its SCV is 0, but the analysis applies more generally (which is important for the overflow bank in the multiclass throttle). In Section 4.1 we observe that the stochastic processes of interest associated with the token bank can be defined in terms of a certain reflection map applied to a net input process. In Section 4.2 we apply this representation to develop a direct Brownian approximation. From this Brownian approximation we obtain renewal-process approximations for the job and token overflow processes, which we apply to analyze the multiclass throttle in Berger and Whitt [1992]. In Section 4.3 and Section 4.4 we develop refinements. Then in Section 4.5 we prove supporting heavy-traffic limit theorems.

4.1 The Two-Sided Regulator

Let $T(t)$ represent the number of tokens in the token bank at time t, and let $O_J(t)$ and $O_T(t)$ represent, respectively, the number of jobs and tokens to overflow in the interval $[0,t]$. It is significant that the stochastic processes $T(t)$, $O_J(t)$, and $O_T(t)$ can be represented *exactly* as the image of the two-sided regulator reflection map on p. 22 of Harrison [1985] applied to the net input stochastic process.

$$N(t) = A_T(t) - A_J(t), \tag{16}$$

where $A_T(t)$ and $A_J(t)$ represent, respectively, the number of tokens and jobs to arrive in $[0,t]$. For the three other related models—the leaky bucket, the G/G/1/C queue, and the two-queue CQN— this representation is not exact, but the error is asymptotically negligible in the heavy-traffic limit. In other words, the token bank model has the same autonomous-service property as the artificial modified system introduced by Borovkov [1965] and applied by Iglehart and Whitt [1970a,b] to prove heavy-traffic limit theorems for the GI/G/C queue (and generalizations).

Harrison [1985] considers stochastic processes with continuous sample paths, but the two-sided regulator is well defined more generally; this follows from Chen and Mandelbaum [1991a,b]. In particular, let $D \equiv D([0, T], R)$ be the space of all right-continuous real-valued functions on $[0, T]$ with left limits everywhere, as in Billingsley [1968]. The two-sided regulator can be defined as the unique map taking $x \in D$ with $0 \le x(0) \le C$ into (z, l, u) in D^3, where $0 \le z(t) \le C$,

$$z(t) = x(t) + l(t) - u(t), \qquad 0 \le t \le T, \tag{17}$$

$l(t)$ and $u(t)$ have nondecreasing sample paths with $l(0) = u(0) = 0$, $l(t)$ increases only when $z(t) = 0$ and $u(t)$ increases only when $z(t) = C$; i.e.,

$$\int_0^T z(t)\, dl(t) = \int_0^T [z(t) - C]\, du(t) = 0. \tag{18}$$

In our application, x, l, u, and z represent sample paths of $N(t)$, $O_J(t)$, $O_T(t)$, and $T(t)$, respectively.

4.2 The Direct Brownian Approximation

The direct Brownian approximation is obtained by simply approximating the net input process $A_T(t)$ - $A_J(t)$ by a Brownian motion (BM) and applying known formulas for the regulated or reflected Brownian motion (RBM) associated with the two-sided regulator, i.e., we use $(Z, L, U) \equiv \{(Z(t), L(t), U(t)) : 0 \le t \le t\}$, where (Z, L, U) is the image of the two-sided regulator applied to a BM process $X \equiv \{X(t) : 0 \le t \le T\}$. We call $L(t)$ and $U(t)$ the *Brownian boundary regulator* processes. (The processes $L(t)$ and $U(t)$ are

also called the *local times* of $Z(t)$ at the boundaries 0 and C, respectively.) Then we have the approximations $T(t) \approx Z(t)$, $O_J(t) \approx L(t)$, and $O_T(t) \approx U(t)$. The resulting formulas are just limits of the M/M/1/C formulas in Section 3, as we will show in Theorem 4.5.

It is natural to choose the drift μ and diffusion coefficient σ^2 of the BM to match the asymptotic behavior of $A_T(t) - A_J(t)$; i.e.,

$$\mu = \lim_{t \to \infty} \frac{E[A_T(t) - A_J(t)]}{t} = r - \lambda \tag{19}$$

and

$$\sigma^2 = \lim_{t \to \infty} \frac{\text{Var}[A_T(t) - A_J(t)]}{t} = rc_r^2 + \lambda c_\lambda^2, \tag{20}$$

assuming that $A_T(t)$ and $A_J(t)$ are independent (which we have assumed), and assuming

$$\frac{EA_T(t)}{t} \to r \quad \text{and} \quad \frac{\text{Var } A_T(t)}{t} \to rc_r^2 \tag{21}$$

and

$$\frac{EA_J(t)}{t} \to \lambda \quad \text{and} \quad \frac{\text{Var } A_J(t)}{t} \to \lambda c_\lambda^2 \tag{22}$$

as $t \to \infty$. Note that (21) and (22) pertain for renewal processes in which the inter-renewal times have means r^{-1} and λ^{-1}, and SCVs c_r^2 and c_λ^2 (but also more generally); see Whitt [1982a, b] for discussion. When the arrival processes are *not* renewal processes, the variability parameters c_r^2 and c_λ^2 in (21) and (22) typically do *not* correspond to the SCVs of stationary intervals between points. The specification of the Brownian motion parameters by (19)–(22) is still appropriate (asymptotically correct) in heavy traffic, but other variability parameters may be more appropriate in lighter traffic. For nonrenewal processes, it is often difficult to choose truly appropriate values for the parameters c_r^2 and c_λ^2; See Sriram and Whitt [1986] and Fendick and Whitt [1989] for further discussion.

Once we have specified the parameters μ and σ^2 in (19) and (20), we obtain the desired approximations from pp. 90–91 of Harrison [1985] or pp. 238–239 of Newell [1982]. Let

$$\theta = \frac{2\mu}{\sigma^2} = \frac{2(\rho - 1)}{\rho c_r^2 + c_\lambda^2}.$$

First, the steady-state number of tokens in the token bank, say $T(\infty)$, has approximately the distribution of $Z(\infty)$ (the limiting distribution of $Z(t)$), which has density

$$p(x) = \begin{cases} \dfrac{1}{C} & \text{if } \rho = 1, \\ \dfrac{\theta e^{\theta x}}{e^{\theta C} - 1} & \text{if } \rho \neq 1. \end{cases} \tag{23}$$

For $\mu < 0$, $p(x)$ in (23) is the density of an exponential distribution with mean $-\theta^{-1} = \sigma^2/2|\mu|$, conditioned on being in the interval $[0, C]$. For $\mu > 0$, $C - Z(\infty)$ has the density of an exponential distribution with mean $\theta^{-1} = \sigma^2/2\mu$, conditioned on being in the interval $[0, C]$.

Of course, the density (23) is inconsistent with the discreteness of the token bank, but it directly yields moments and it can be discretized to yield an approximating probability mass function, if desired. In particular, a natural candidate for an approximating probability mass function is truncated geometric distribution, which is the discrete analog of the truncated exponential distribution in (23). For example, when $\rho < 1$ ($\mu < 0$), we would take the geometric distribution on the nonnegative integers with mean $-\theta^{-1}$, conditioned on being less than or equal to C. Of course, for $\rho = 1$ ($\mu = 1$), we would use the discrete uniform distribution on the set $\{0, 1, \ldots, C\}$. For further discussion about discretization, see Kimura [1985] and Section 8.5.

We obtain the approximation for the steady-state mean number of tokens in the bank directly from (23); it is

$$ET(\infty) \approx EZ(\infty) = \begin{cases} \dfrac{C}{2} & \text{if } \rho = 1, \\ \dfrac{C}{1 - e^{-\theta C}} - \dfrac{1}{\theta} & \text{if } \rho \neq 1. \end{cases} \tag{24}$$

Next, the approximate overflow rates are

$$\lambda' \approx \alpha \equiv \lim_{t \to \infty} \frac{L(t)}{t} = \begin{cases} \dfrac{\sigma^2}{2C} & \text{if } \rho = 1, \\ \dfrac{\mu}{e^{\theta C} - 1} & \text{if } \rho \neq 1, \end{cases} \tag{25}$$

and

$$r' \approx \beta \equiv \lim_{t \to \infty} \frac{U(t)}{t} = \begin{cases} \dfrac{\sigma^2}{2C} & \text{if } \rho = 1, \\ \dfrac{\mu}{1 - e^{-\theta C}} & \text{if } \rho \neq 1, \end{cases} \tag{26}$$

If we let $X(0) = Z(0) \stackrel{d}{=} Z(\infty)$, where $\stackrel{d}{=}$ denotes equality in distribution, then we also have $\alpha = EL(1)$ in (25) and $\beta = EU(1)$ in (26). Note that the overflow rates in (25) and (26) satisfy the conservation law in (2). From (25) and (26) we see that the approximating overflow rates λ' and r' are continuous functions of the parameters μ, σ^2, and C. Moreover, the overflow rates λ' and r' are increasing functions of σ^2 and decreasing functions of C. For further discussions about the implications of (25), see Section 9.

It is significant that the blocking probabilities λ'/λ and r'/r, which are obtained from (25) and (26), do not equal the values of the diffusion density at the boundaries. Rather,

$$\begin{aligned} \frac{\lambda'}{\lambda} &= \left(\frac{\sigma^2}{2\lambda} \right) p(0) = \left(\frac{\rho c_r^2 + c_\lambda^2}{2} \right) p(0), \\ \frac{r'}{r} &= \left(\frac{\sigma^2}{2r} \right) p(C) = \left(\frac{c_r^2 + \rho^{-1} c_\lambda^2}{2} \right) p(C). \end{aligned} \tag{27}$$

Moreover, we see from (23) that when $\rho = 1$ the density $p(x)$ is insensitive to the SCVs c_r^2 and c_λ^2, whereas from (25)–(27) we see that the blocking probabilities are not. (Most previous heuristic diffusion approximations for blocking in G/G/1/C queues have not exploited (25) and (26); see Section 6 of Coffmann and Reiman [1984] for further discussion.)

From the regenerative analysis on p. 86 of Harrison [1985], we see that $L(t)$ and $U(t)$ also obey central limit theorems as $t \to \infty$, and

$$\lim_{t \to \infty} t^{-1} \operatorname{Var} L(t) = \sigma_L^2 \quad \text{and} \quad \lim_{t \to \infty} t^{-1} \operatorname{Var} U(t) = \sigma_U^2.$$

We apply these asymptotic variance constants σ_L^2 and σ_U^2 to develop renewal-process approximations for the overflow processes. In particular, we approximate the overflow SCVs c_J^2 and c_T^2 by

$$c_J^2 \approx \alpha^{-1} \sigma_L^2 \quad \text{and} \quad c_T^2 \approx \beta^{-1} \sigma_U^2. \tag{28}$$

(The use of α and β in (28) is explained by (21) and (22).) At the end of Section 4.5, we apply Theorem 3.1 and heavy-traffic limit theorems to determine the formulas for σ_L^2 and σ_U^2 in (29) below. A direct proof has been provided by Williams [1991].

THEOREM 4.1. For RBM with barriers at 0 and C, drift coefficient μ, diffusion coefficient σ^2, and $\theta = 2\mu/\sigma^2$,

$$\alpha^{-1} \sigma_L^2 = \beta^{-1} \sigma_U^2 = \begin{cases} \dfrac{2C}{3} & \text{if } \mu = 0, \\ \dfrac{2(1 - e^{2\theta C}) + 4\theta C e^{\theta C}}{-\theta(1 - e^{\theta C})^2} & \text{if } \mu \neq 0. \end{cases} \tag{29}$$

As in Theorem 3.1, we have $\alpha^{-1}\,\sigma_L^2 = \beta^{-1}\,\sigma_U^2$. Note that $\alpha^{-1}\,\sigma_L^2$ is an *even function* of θ; i.e., $(\alpha^{-1}\sigma_L^2)(-\theta) = (\alpha^{-1}\,\sigma_L^2)(\theta)$. Note that $\alpha^{-1}\,\sigma_L^2$ is independent of σ_1^2 for $\mu = 0$, but not for $\mu \neq 0$. For $\mu \neq 0$, $\alpha^{-1}\,\sigma_L^2$ is *decreasing* in $|\theta|$ with $(\alpha^{-1}\,\sigma_L^2)(\theta) \to (\alpha^{-1}\,\sigma_L^2)(0)$ as $\theta \to 0$. Also note that provided C is not too small, the Brownian approximation yields $\alpha^{-1}\,\sigma_L^2 \approx -2/\theta$ when $r \ll \lambda$ and yields $\alpha^{-1}\,\sigma_L^2 \approx 2/\theta$ when $\lambda \ll r$. Thus, since $\theta(\rho) \to -2/c_\lambda^2$ as $\rho \to 0$ and $\theta(\rho) \to 2/c_r^2$ as $\rho \to \infty$, the Brownian approximation yields $c_J^2 \approx c_\lambda^2$ when $r \ll \lambda$ and yields $c_J^2 \approx c_r^2$ when $\lambda \ll r$, consistent with intuition. See Section 6 for further discussion.

Combining (28) and (29), we obtain the Brownian approximation for c_J^2 and c_T^2. If we want a full renewal process as an approximation to the job overflow process, we fit an inter-renewal distribution on the positive real line to the mean $1/\lambda'$ for λ' in (25) and the SCV c_J^2 in (28); e.g., see Whitt [1982a]. To further match the character of the true job overflow process, we suggest applying the model structure to obtain additional properties of the overflow stream. For example, suppose that the job arrival process is a renewal process with an interarrival-time distribution that is the mixture of two distributions, one with small mean and the other with large mean, as with the H_2 distribution and the bursty on-off renewal process in Sriram And Whitt [1986]. Then we propose letting the approximating interoverflow distribution be the mixture of the component distribution in the mixture with the small mean and an exponential distribution, typically with a much larger mean. Alternatively, both components of the mixture in the approximating interoverflow distribution could be exponential; then the analysis above would only determine the mean of one component exponential distribution in the approximation. If the overall arrival process is a Poisson process, then we would just use the exponential interarrival-time distribution as one component in the approximating interoverflow distribution. In either case, the overall mean $(\lambda')^{-1}$ and SCV c_J^2 determine the remaining two parameters: the mixing probability and the mean of the additional exponential distribution. This approach exploits the two parameters λ' and c_J^2 as well as properties of the interarrival-time distribution; see Whitt [1982a, 1989] for a discussion of fitting and improved approximations.

4.3 The M/M/1/C Refinement

The formulas in (23)–(29) are extremely appealing for their relative simplicity, but we can obtain improved numerical accuracy in some cases by making refinements. First, it is reasonable to require that the approximation be exact for the M/M/1/C case in Section 3. A simple way to achieve this is to multiply the M/M/1/C values in Section 3 by the ratio of the diffusion values. In doing so we treat the cases $\rho < 1$ and $\rho > 1$ differently.

Let $ET(\infty;\, \rho,\, c_r^2,\, c_\lambda^2)$ be the steady-state mean number of tokens as a function of the parameters ρ, c_r^2, and c_λ^2; let $ET_M(\infty;\, \rho)$ be the corresponding mean in the M/M/1/C model with the same ρ; and let $EZ(\infty;\, \rho,\, c_r^2,\, c_\lambda^2)$ be the diffusion approximation in (24) with $\theta = 2(\rho - 1)/(\rho c_r^2 + c_\lambda^2)$. then the suggested refinement is

$$ET(\infty;\, \rho,\, c_r^2,\, c_\lambda^2) = \begin{cases} ET_M(\infty;\, \rho)\,\dfrac{EZ(\infty;\, \rho,\, c_r^2;\, c_\lambda^2)}{EZ(\infty;\, \rho,\, 1,\, 1)} & \text{if } \rho \le 1, \\ C - ET_M(\infty,\, \rho^{-1})\,\dfrac{EZ(\infty;\, \rho^{-1},\, c_\lambda^2,\, c_r^2)}{EZ(\infty;\, \rho^{-1},\, 1,\, 1)} & \text{if } \rho > 1. \end{cases} \tag{30}$$

The idea in treating the case $\rho > 1$ differently is to make the adjustment to smaller numbers, so that we make smaller absolute errors. Note that we switch the roles of c_λ^2 and c_r^2 when $\rho > 1$ in (30). When $\rho > 1$, we are looking at the dual queue, which is C minus the number of tokens, so that the job and token arrival processes switch roles. Also notice that $\theta(\rho^{-1}, c_\lambda^2, c_r^2) = -\theta(\rho, c_r^2, c_\lambda^2)$.

Similarly, let $\lambda'(r, \lambda, c_r^2, c_\lambda^2)$ be the job overflow rate as a function of the parameters, let $\lambda'_M(r, \lambda)$ be the M/M/1/C formula in (5) and let $\lambda'_D(r, \lambda, c_r^2, c_\lambda^2)$ be the diffusion formula in (25). Define corresponding quantities for the tokens. Then suggested refined approximation is

$$\begin{aligned} r'(r, \lambda, c_r^2, c_\lambda^2) &= r'_M(r, \lambda) \frac{r'_D(r, \lambda, c_r^2, c_\lambda^2)}{r'_D(r, \lambda, 1, 1)} \\ &= r \left(\frac{1 - \rho^{-1}}{1 - \rho^{-(C+1)}} \right) \left(\frac{e^{2(1-\rho)C/(1+\rho)} - 1}{e^{2(1-\rho)C/(c_\lambda^2 + \rho c_r^2)} - 1} \right) \end{aligned} \tag{31}$$

and

$$\lambda' = r' + \lambda - r \quad \text{if } \rho < 1;$$

$$\begin{aligned} \lambda'(r, \lambda, c_r^2, c_\lambda^2) &= \lambda'_M(r, \lambda) \frac{\lambda'_D(r, \lambda, c_r^2, c_\lambda^2)}{\lambda'_D(r, \lambda, 1, 1)} \\ &= \lambda \left(\frac{\rho - 1}{\rho^{C+1} - 1} \right) \left(\frac{e^{2(\rho-1)C/(1+\rho)} - 1}{e^{2(\rho-1)C/(c_\lambda^2 + \rho c_r^2)} - 1} \right) \end{aligned} \tag{32}$$

and

$$r' = \lambda' + r - \lambda \quad \text{if } \rho > 1;$$

and

$$\begin{aligned} \lambda'(r, \lambda, c_r^2, c_\lambda^2) = r'(r, \lambda, c_r^2, c_\lambda^2) &= \lambda'_M(\lambda, \lambda) \frac{\lambda'_D(\lambda, \lambda, c_r^2, c_\lambda^2)}{\lambda'_D(\lambda, \lambda, 1, 1)} \\ &= \frac{\lambda(c_r^2 + c_\lambda^2)}{2(C + 1)} \quad \text{if } \rho = 1. \end{aligned} \tag{33}$$

Note that we should have $\lambda' < r'$ when $\rho > 1$ and $r' < \lambda'$ when $\rho < 1$, so that we apply the multiplicative adjustment to what we anticipate will be the smaller rate in (31) and (32), and then apply the conservation law to determine the other rate. Since $\theta(r, \lambda, c_r^2, c_\lambda^2) = -\theta(\lambda, r, c_\lambda^2, c_r^2)$, $\lambda'_M(r, \lambda) = r'_M(\lambda, r)$, and $\lambda'_D(r, \lambda, c_r^2, c_\lambda^2) = r'(\lambda, r, c_\lambda^2, c_r^2)$. Hence, there is no need to use the dual queue representation in (32).

Finally, we define a refined approximation for $c_J^2(\rho, c_r^2, c_\lambda^2)$ in the same way, drawing on (6), (28), and (29). Let $c_{JM}^2(\rho)$ be the M/M/1/C formula in (6) and let $c_{JD}^2(\rho, c_\lambda^2, c_r^2)$ be the diffusion formula in (28) and (29). The suggested refinement is

$$c_J^2(\rho, c_r^2, c_\lambda^2) = c_T^2(\rho, c_r^2, c_\lambda^2) = c_{JM}^2(\rho)\,\frac{c_{JD}^2(\rho, c_r^2, c_\lambda^2)}{c_{JD}^2(\rho, 1, 1)}. \tag{34}$$

Since $c_{JM}^2 = c_{TM}^2$ and $c_{JD}^2 = c_{TD}^2$, there is no need to consider the dual queue.

4.4 The GI/G/1/∞ Refinement

In this section we incorporate refinements previously developed for single-server queues with unlimited waiting space. The main idea is to let $-\theta^{-1}$ in the Brownian approximation (in (23)–(26) and (29)) be the steady-state mean number of tokens in the token bank when $C = \infty$ in the case $\rho < 1$. With this refinement, we can use any algorithm or approximation formula for the GI/G/1/∞ queue. Below we specify a specific procedure using only the parameters ρ, c_r^2, and c_λ^2, but it would also be natural to incorporate additional information about the arrival processes if it is available. (In Section 4.2 we indicated how to obtain additional information for the overflow processes.) Some ways to do this are in Whitt [1989], Fendick and Whitt [1989], and references cited there.

Drawing on Kraemer and Langenbach-Belz [1976] and p. 17 of Whitt [1985], we approximate the steady-state mean in the GI/G/1/∞ model with interarrival-time mean r^{-1} and SCV c_r^2, service time mean λ^{-1} and SCV c_λ^2, and $\rho = r/\lambda < 1$ by

$$m \equiv m(\rho, c_r^2, c_\lambda^2) \approx \rho + \frac{\rho^2(c_r^2 + c_\lambda^2)}{2(1-\rho)}\, g(\rho, c_r^2, c_\lambda^2), \tag{35}$$

where

$$g(\rho, c_r^2, c_\lambda^2) = \begin{cases} \exp\left[-\dfrac{2(1-\rho)}{3\rho}\,\dfrac{(1-c_r^2)^2}{c_r^2 + c_\lambda^2} \right], & c_r^2 \le 1, \\ \exp\left[-(1-\rho)\,\dfrac{c_r^2 - 1}{c_r^2 + 4c_\lambda^2} \right], & c_r^2 > 1. \end{cases} \tag{36}$$

The function g is always less than or equal to 1. Note that g=1 when $c_r^2 = 1$, so that for the M/G/1/∞ queue (35) is exact. Indeed (35) with g eliminated is a reasonable approximation. The correction factor g reflects the fact that m depends not just on the sum $c_r^2 + c_\lambda^2$ but also on the location of the variability. For example, the mean m in a D/M/1/∞ model is less than in an M/D/1/∞ model with the same ρ. Thus, it is significant that $g(\rho, c_r^2, c_\lambda^2)$ is not symmetric in c_λ^2 and c_r^2.

There are three cases for the approximation: $\rho < 1$, $\rho = 1$, and $\rho > 1$. When $\rho < 1$, we let $\theta = -m^{-1}$ for $m \equiv m(\rho, c_r^2, c_\lambda^2)$ in (35); then we use the approximations in Section 4.3. The approximations in (30), (31), and (34) are changed only by having the

diffusion quantities computed in terms of a new θ. Note that θ must be recomputed when the SCVs are both 1 as well as when the SCVs are c_r^2 and c_λ^2.

For example, for the mean, we use (30) with $EZ(\infty; \rho, c_r^2, c_\lambda^2)$ and $EZ(\infty; \rho, 1, 1)$ being given by (24). However, $EZ(\infty; \rho, c_r^2, c_\lambda^2)$ is now based on $\theta \equiv \theta(\rho, c_r^2, c_\lambda^2) = -1/m(\rho, c_r^2, c_\lambda^2)$, where $m \equiv m(\rho, c_r^2, c_\lambda^2)$ is given by (35) in terms of the parameters ρ, c_r^2, and c_λ^2, whereas $EZ(\infty; \rho, 1, 1)$ is based on $\theta \equiv \theta(\rho, 1, 1) = -1/m(\rho, 1, 1)$, where $m \equiv m(\rho, 1, 1)$ is given by (35) in terms of the parameters ρ, 1, and 1.

There is no change from Section 4.3 when $\rho = 1$. We set $\theta = 0$ just as before. When $\rho > 1$, we exploit the GI/G/1/∞ refinement by working with the dual queue. We switch the roles of the token and job arrival processes, and look at the number of empty spaces in the token bank. To perform the approximation, we first replace ρ by ρ^{-1}, c_r^2 by c_λ^2, and c_λ^2 by c_r^2. Then we apply the GI/G/1/∞ and M/M/1/C refinements above to calculate the performance measures for the dual queue. The mean in the token bank is then C minus the mean in the dual queue; the job overflow rate λ' in the token bank is the token overflow rate r' in the dual queue; the token overflow rate r' in the token bank is the job overflow rate λ' in the dual queue; and the overflow SCVs c_T^2 and c_J^2 in the token bank are the same as the overflow SCVs in the dual queue (which are equal).

4.5 Supporting Limit Theorems

To provide additional support for the Brownian approximation above, we prove some heavy-traffic limit theorems. For our first heavy-traffic limit theorem, it is significant that the two-sided regulator is a continuous map. This was observed without proof in Section 6.8 of Whitt [1969] and in Kennedy [1973]. The continuity was proved explicitly by Chen and Mandelbaum [1991a,b]; the two-sided regulator arises in the special case of the two-queue closed network. However, we present a different proof here for this special case, which we believe is of interest. We actually establish a stronger Lipschitz property for z, which provides a basis for establishing forms of model stability and rates of convergence; see Whitt [1974], Dupuis and Ishii [1991], and Kalashnikov and Rachev [1990], Chen and Whitt [1991], and Theorem 4.4.

Below let a common subscript index the associated element of the function space D (defined in Section 4.1), e.g., (z_i, l_i, u_i) is the image of x_i under the two-sided regulator. Let $\|\cdot\|$ be the supremum norm, i.e., $\|x\| = \sup_{0 \le t \le T} |x(t)|$.

THEOREM 4.2. (a) $\|z_1 - z_2\| \le 2\|x_1 - x_2\|$.

(b) If $\|x_n - x\| \to 0$ as $n \to \infty$, then $\|l_n - l\| \to 0$ and $\|u_n - u\| \to 0$ as $n \to \infty$.

Proof. (a) Any function in D can be approximated arbitrarily closely in the supremum norm by a piecewise-constant function with finitely many discontinuities; see p. 110 of Billingsley [1968]. Hence, it suffices to let x_1 and x_2 be such piecewise-constant functions. We carry out the proof by mathematical induction over the successive times at which at least one of these functions has a jump. Let t_n be the n^{th} jump epoch. Suppose that $\|x_1 - x_2\| = \epsilon$. Let $\Delta_n = x_1(t_n) - x_2(t_n)$ and $\Gamma_n = z_1(t_n) - z_2(t_n)$. We are given that $|\Delta_n| \le \epsilon$ for all n. By induction, we establish that

$$\Delta_n - \epsilon \le \Gamma_n \le \Delta_n + \epsilon \quad \text{for all } n, \tag{37}$$

from which the desired conclusion follows. Note that $z_i(0) = x_i(0)$, so that $\Gamma_0 = \Delta_0$ and (37) holds for $n = 0$. Suppose that (37) holds for all $k \le n$. Then, by considering the possible jumps at t_{n+1}, we see that

$$\Gamma_{n+1} \le \begin{cases} 0 \le \Delta_{n+1} + \epsilon & \text{if } z_2(t_{n+1}) = C \text{ or } z_1(t_{n+1}) = 0, \\ \Gamma_n + \Delta_{n+1} - \Delta_n \le \Delta_{n+1} + \epsilon & \text{otherwise,} \end{cases}$$

and

$$\Gamma_{n+1} \ge \begin{cases} 0 \ge \Delta_{n+1} - \epsilon & \text{if } z_2(t_{n+1}) = 0 \text{ or } z_1(t_{n+1}) = C, \\ \Gamma_n + \Delta_{n+1} - \Delta_n \ge \Delta_{n+1} - \epsilon & \text{otherwise,} \end{cases}$$

(b) Suppose that $\|x_n - x\| \to 0$. As in part (a), it suffices to let x be piecewise constant with finitely many discontinuities in $[0, T]$. Let z, l, and u be the image of x under the reflection map. It is easy to see that these functions are also piecewise constant with discontinuities only at discontinuity points of x. Let t_k be the k^{th} discontinuity point of x. It suffices to restrict attention to those subintervals $[t_k, t_{k+1})$ for which z is at a boundary, i.e., for which $z(t_k) = 0$ or $z(t_k) = C$, because by part (a), $\|z_n - z\| \to 0$ as $n \to \infty$; i.e., for the given z, we can choose n_0 so that $0 < z_n(t) < C$ for all $n \ge n_0$ and all other t. Hence, l, u, l_n, and u_n with $n \ge n_0$ are constant everywhere except possibly over these subintervals $[t_k, t_{k+1})$. Moreover, over each of these subintervals it suffices to consider only the one barrier z is at. We then use the well-known continuity of the one-sided regulator reflection map over each of these intervals. For example, when there is no upper barrier, $z = x + l$ by (17) and $\|l_1 - l_2\| \le \|z_1 - z_2\| + \|x_1 - x_2\|$. ■

EXAMPLE 4.1. To see that the bound in Theorem 4.2(a) is sharp, let $C \ge 2$, $x_1(t) = z_1(t) = 0$, and $x_2(t) = -I_{[1,2)}(t) + I_{[2,3)}(t)$, $0 \le t \le 3 = T$, where $I_A(t)$ is the indicator function of the set A. Then $z_2(t) = 2I_{[2,3)}(t)$, $0 \le t \le T$, $\|x_1 - x_2\| = 1$, and $\|z_1 - z_2\| = 2$.

Let d be the metric from p. 111 of Billingsley [1968] inducing the Skorohod J_1 topology on the space D. The properties of Theorem 4.2 carry over to (D, d), as is shown by Proposition 2.4 of Chen and Whitt [1991].

COROLLARY. (a) $d(z_1, z_2) \le 2d(x_1, x_2)$.
(b) If $d(x_n, x) \to 0$ as $n \to \infty$, then $d(l_n, l) \to 0$ and $d(u_n, u) \to 0$ as $n \to \infty$.

EXAMPLE 4.2. It is interesting that the maps from x to l and u are actually *not* Lipschitz. For x_i of bounded variation (as in the token bank), $x_i = y_i^{\uparrow} + y_i^{\downarrow}$, where $y_i^{\uparrow}$ is nondecreasing and $y_i^{\downarrow}$ is nonincreasing. We might hope to bound $\|l_1 - l_2\|$ and $\|u_1 - u_2\|$ by

$\|y_1^\dagger - y_2^\dagger\| + \|y_1^\downarrow - y_2^\downarrow\|$, but this is not possible either. To see this, suppose that $C > \epsilon$ and let

$$y_1^\dagger(t) = \sum_{k=0}^{n-1} [CI_{[2kT/2n,T]}(t) + (C + \epsilon)I_{[(2k+1)T/2n,T]}(t)], \tag{38}$$

$$y_2^\dagger(t) = \sum_{k=0}^{n-1} [(C + \epsilon)I_{[2kT/2n,T]}(t) + CI_{[(2k+1)T/2n,T]}(t)], \tag{39}$$

and

$$y_1^\downarrow(t) = y_2^\downarrow(t) = -\sum_{k=0}^{n-1} (2C + \epsilon)CI_{[(2k+1)T/2n,T]}(t), \qquad 0 \le t \le T. \tag{40}$$

Then $\|y_1^\dagger - y_2^\dagger\| = \epsilon$, $\|x_1 - x_2\| = \epsilon$, $\|z_1 - z_2\| = 0$, and $l_1(t) = 0$ for all t, but

$$l_2(t) = \sum_{k=0}^{n-1} \epsilon I_{[(2k+1)T/2n,T]}(t), \tag{41}$$

so that

$$d(l_1, l_2) = \|l_1 - l_2\| = \|l_2\| = n\epsilon = n\|x_1 - x_2\| = nd(x_1, x_2).$$

Theorem 4.2 provides the basis for a heavy-traffic limit theorem, with convergence to reflected Brownian motion (RBM), where the reflection is by the two-sided regulator. For this purpose, we consider a sequence of token banks indexed by n. The n^{th} system has capacity C_n and job and token arrival processes $A_{Jn} \equiv \{A_{Jn}(t): t \ge 0\}$ and $A_{Tn} \equiv \{A_{Tn}(t): t \ge 0\}$. Let $\mathbf{A}_{Jn}$ and $\mathbf{A}_{Tn}$ be associated normalized processes, defined by

$$\mathbf{A}_{Jn}(t) = \frac{A_{Jn}(nt) - \lambda_n nt}{\sqrt{n}}, \qquad t \ge 0, \tag{42}$$

and

$$\mathbf{A}_{Tn}(t) = \frac{A_{Tn}(nt) - r_n nt}{\sqrt{n}}, \qquad t \ge 0. \tag{43}$$

This is the typical normalization associated with the heavy-traffic conditions specified below. Let $\Rightarrow$ denote convergence is distribution.

For the limit theorem, we make the following assumption, which is satisfied when A_{Jn} and A_{Tn} are independent renewal processes with uniformly bounded third moments, but also more generally.

HEAVY-TRAFFIC ASSUMPTION. Suppose that $\lambda_n \to \lambda$, $0 < \lambda < \infty$, $r_n \to r$, and $\sqrt{n}(r_n - \lambda_n) \to \mu$ as $n \to \infty$ (so that $\lambda = r$); $C_n/\sqrt{n} \to C$, $0 < C < \infty$, as $n \to \infty$; and $(\mathbf{A}_{Tn}, \mathbf{A}_{Jn}) \Rightarrow (\sqrt{rc_r^2}\, B_T, \sqrt{\lambda c_\lambda^2}\, B_J)$ in D^2 as $n \to \infty$, where B_T and B_J are independent standard (zero drift, unit variance) Brownian motions.

To express the limit theorem, consider the process (T_n, O_{Jn}, O_{Tn}) defined by the two-sided regulator applied to $(A_{Tn} - A_{Jn})$. Then consider the associated normalized processes $\mathbf{T}_n$, $\mathbf{O}_{Jn}$, and $\mathbf{O}_{Tn}$, defined by

$$\mathbf{T}_n(t) = \frac{T_n(nt)}{\sqrt{n}}, \quad \mathbf{O}_{Jn}(t) = \frac{O_{Jn}(nt)}{\sqrt{n}}, \quad \text{and} \quad \mathbf{O}_{Tn}(t) = \frac{O_{Tn}(nt)}{\sqrt{n}}. \tag{44}$$

Note that $(\mathbf{T}_n, \mathbf{O}_{Jn}, \mathbf{O}_{Tn})$ is the image of the $(\mathbf{A}_{Tn} - \mathbf{A}_{Jn})$ under the two-sided regulator for each n. Theorem 4.2 and the continuous mapping theorem, Theorem 5.1 of Billingsley [1968], imply the following (known) result.

THEOREM 4.3. Under the heavy-traffic assumption above

$$(\mathbf{T}_n, \mathbf{O}_{Jn}, \mathbf{O}_{Tn}) \Rightarrow (Z, L, U) \text{ in } D^3 \qquad \text{as } n \to \infty,$$

where (Z, L, U) is the image of the two sided regulator with barriers at 0 and C applied to $X \equiv \sqrt{\lambda(c_r^2 + c_\lambda^2)}\, B + \mu e$, B is a standard Brownian motion, and e is the identity map on $[0, T]$; i.e., X is a BM with drift μ and diffusion coefficient $\sigma^2 \equiv \sqrt{\lambda(c_r^2 + c_\lambda^2)}$.

Proof. Let

$$\mathbf{X}_n(t) = \frac{A_{Tn}(nt) - A_{Jn}(nt)}{\sqrt{n}}, \quad t \geq 0. \tag{45}$$

By the heavy-traffic assumption and the continuous mapping theorem applied with subtraction (which is continuous at functions with continuous paths, see Section 4 of Whitt [1980]),

$$\mathbf{X}_n - \sqrt{n}(r_n - \lambda_n)e \Rightarrow \sqrt{rc_r^2}\, B_T - \sqrt{\lambda c_\lambda^2}\, B_J \stackrel{d}{=} \sqrt{rc_r^2 + \lambda c_\lambda^2}\, B \text{ in } D \qquad \text{as } n \to \infty,$$

where $e(t) = t$. Since $\sqrt{n}(r_n - \lambda_n)e \to \mu e$ in D, $\mathbf{X}_n \Rightarrow X$. Finally apply the continuous mapping theorem again with Corollary to Theorem 4.2. ∎

Theorem 4.3 implies a limit for the proportions of jobs and token blocked. Let $\Lambda_n'(T)$ and $R_n'(T)$ be the proportions of jobs and tokens blocked over the interval $[0, nT]$ in the n^{th} model; i.e.,

$$\Lambda_n'(T) = \frac{O_{Jn}(nT)}{nT} \quad \text{and} \quad R_n'(T) = \frac{O_{Tn}(nT)}{nT}. \tag{46}$$

COROLLARY. Under the heavy-traffic assumption,

$$\sqrt{n}\,\Lambda_n'(T) \Rightarrow \frac{L(T)}{T} \quad \text{and} \quad \sqrt{n}\,R_n'(T) \Rightarrow \frac{U(T)}{T} \quad \text{as } n \to \infty$$

for each T.

Harrison [1985] has shown that $T^{-1}L(T) \to \alpha$ in (25) and $T^{-1}U(T) \to \beta$ in (26) as $T \to \infty$. By similar regenerative arguments, $\sqrt{n}\Lambda_n'(T) \to \sqrt{n}\lambda_n'$ and $\sqrt{n}R_n'(T) \to \sqrt{n}r_n'$ as $T \to \infty$ if the job and token arrival processes are independent renewal processes, at least one of whose interarrival-time distributions is phase type. We *conjecture* that

$$\sqrt{n}\lambda_n' \to \alpha \quad \text{and} \quad \sqrt{n}r_n'\,\beta \quad \text{as } n \to \infty \tag{47}$$

if, in addition, the job and token interarrival times have uniformly bounded third moments and the heavy-traffic assumption is satisfied; i.e., we conjecture that the iterated limit in which first $T \to \infty$ and then $n \to \infty$ is valid, but we have not yet established it. If we let $X_n(0)$ have the distribution of $T_n(\infty)$ and if $X_n(0) \Rightarrow Z(\infty)$ as $n \to \infty$, then $\lambda_n' = E\Lambda_n'(T)$ and $r_n' = ER_n'(T)$, so that the desired result would follow from the corollary to Theorem 4.3 by uniform integrability. However, there are gaps here as well. Nevertheless, the corollary provides additional support for the approximation in (25)–(26) and (31)–(33). In Theorem 4.5 we establish the desired limit for the long-run blocking probabilities in the M/M/1/C special case.

We can apply part (a) of the corollary to Theorem 4.2 plus strong approximation results in the literature to obtain a bound on the rate of convergence of $\mathbf{T}_n$ to Z. For this purpose, let m be the Prohorov metric on the space of probability measures on D; see p. 238 of Billingsley [1968]. When we use random elements as arguments of m, we mean their probability laws.

THEOREM 4.4. Suppose that the token and job arrival processes are renewal processes with interarrival-time random variables U_n and V_n in the n^{th} model. Let $\sqrt{n}(r_n - \lambda_n) = \mu$ for each n.

(a) if $E(e^{tU_n}) < \infty$ and $E(e^{tV_n}) < \infty$ for t in a neighborhood of the origin, then

$$m(\mathbf{T}_n, Z) \le K\frac{\log n}{\sqrt{n}}$$

for some constant K.

(b) If $E\,U_n^r < \infty$ and $E\,V_n^r < \infty$ for some $r < 2$, then

$$m(\mathbf{T}_n, Z) \leq Kn^{-(r-2)/2(r+1)}$$

for some constant K.

Proof. The condition in (a) implies that

$$m(\mathbf{A}_{Jn}, \sqrt{\lambda c_\lambda^2}\, B_J) \leq K_1 \frac{\log n}{\sqrt{n}} \quad \text{and} \quad m(\mathbf{A}_{Tn}, \sqrt{rc_r^2}\, B_T) \leq \frac{K_2 \log n}{\sqrt{n}}$$

for some constants K_1 and K_2, while the condition in (b) implies that

$$m(\mathbf{A}_{Jn}, \sqrt{\lambda c_\lambda^2}\, B_J) \leq K_1 n^{-(r-2)/2(r+1)} \quad \text{and} \quad m(\mathbf{A}_{Tn}, \sqrt{rc_r^2}\, B_T) \leq K_2 n^{-(r-2)/2(r+1)};$$

see Corollary 4.1 of Csörgő, Horváth, and Steinebach [1987]. We only discuss (a) because the remaining argument for (b) is the same. We apply the triangle inequality to get

$$m(\mathbf{X}_n - \sqrt{n}(r_n - \lambda_n)e, \sqrt{rc_r^2 + \lambda c_\lambda^2}\, B) \leq (K_1 + K_2) \frac{\log n}{\sqrt{n}},$$

and once more to get

$$m(\mathbf{X}_n, \mathbf{X}) \leq (K_1 + K_2) \frac{\log n}{\sqrt{n}} + | \sqrt{n}(r_n - \lambda_n) - \mu |.$$

But, by assumption, $\sqrt{n}(r_n - \lambda_n) - \mu = 0$. We conclude by applying part (a) of the corollary to Theorem 4.2 together with Theorem 3.2 of Whitt [1974], which shows that

$$m(f(X_1), f(X_2)) \leq \max \{1, K\}\, m(X_1, X_2)$$

when f is Lipschitz with modulus K. ■

In order to prove Theorem 4.1 in Section 4.2, and for its own sake, we now consider the heavy-traffic behavior of the steady-state characteristics of the M/M/1/C model in Section 3. The appropriate limiting regime is indicated by the heavy-traffic assumption above, but now we apply the explicit steady-state formulas from Section 3. Let $[x]$ be the greatest integer less than or equal to x.

THEOREM 4.5. Consider a sequence of M/M/1/C models indexed by n with $\lambda_n = \lambda = 1$ and $r_n = \rho_n$ for all n. If $\sqrt{n}(\rho_n - 1) \to \mu$ and $C_n/\sqrt{n} \to C$, then

$$\sqrt{n} p_n([x\sqrt{n}]) \to p(x) \qquad \text{for each } x, \tag{48}$$

$$\frac{E[T_n(\infty)]}{\sqrt{n}} \to E[Z(\infty)], \tag{49}$$

$$\sqrt{n}\lambda_n' \to \alpha, \qquad \sqrt{n}r_n' \to \beta, \tag{50}$$

and

$$\frac{c_{Jn}^2}{\sqrt{n}} = \frac{c_{Tn}^2}{\sqrt{n}} \to \alpha^{-1}\,\sigma_L^2 \qquad \text{as } n \to \infty, \tag{51}$$

where the limits are defined in (23)–(29) with $\sigma^2 = 2$ and $\theta = \mu$.

Proof. Apply (3)–(6) and the fact that

$$\rho_n^{\sqrt{n}} = (1 + (\rho_n - 1))^{(\rho_n - 1)^{-1}\sqrt{n}(\rho_n - 1)} \to e^{\mu} \qquad \text{as } n \to \infty,$$

because $(1 + (\rho_n - 1))^{(\rho_n - 1)^{-1}} \to e$ and $\sqrt{n}(\rho_n - 1) \to \mu$ as $n \to \infty$. Note that $a_n^{b_n} \to a^b$ when $a_n \to a$ and $b_n \to b$ with $0 < a < \infty$ and $-\infty < b < \infty$. It is easy to apply this approach directly for the case $\mu \neq 0$. For $\mu \neq 0$, it is convenient to apply the results for $\mu \neq 0$ together with the continuity of the RBM quantities in μ. For example, let $p_n'(x)$ be the state probabilities associated with ρ_n' for which $\rho_n' < \rho_n$ and $\sqrt{n}(\rho_n' - 1) \to \mu' < 0$ as $n \to \infty$. Similarly, let $p_n''(x)$ be the state probabilities associated with ρ_n'' for which $\rho_n'' > \rho_n$ and $\sqrt{n}(\rho_n'' - 1) \to \mu > 0$ as $n \to \infty$. Then the normalized state probabilities of interest $\sqrt{n}p_n([x\sqrt{n}])$ are bounded above and below by

$$\min\{\sqrt{n}p_n'(C_n), \sqrt{n}p_n''(0)\} \le \sqrt{n}p_n([x\sqrt{n}]) \le \max\{\sqrt{n}p_n'(0), \sqrt{n}p_n''(C_n)\}.$$

Moreover, the four bounding terms all converge as $n \to \infty$ and in turn these limits converge to $1/C$ as $\mu' \uparrow 0$ and $\mu'' \downarrow 0$. Hence, $\sqrt{n}p_n([x\sqrt{n}]) \to 1/C$ as $n \to \infty$ when $\mu = 0$. Similar reasoning applies in the other cases. ■

We now apply Theorem 4.5 to prove Theorem 4.1, modulo two technical gaps. A direct proof using properties of RBM has been provided by Williams [1991]. Nevertheless, the limiting argument below seems to be of considerable interest.

Partial proof of Theorem 4.1. We deduce these properties of RBM by considering the limiting behavior of the M/M/1/C model. (The same can be done for (23)–(26).) First, for any given C, μ and σ^2, choose C_n, λ, λ_n, and r_n so that $C_n = \sqrt{n}\,C$, $\sqrt{n}(r_n - \lambda_n) \to \mu$, and $2\lambda = \sigma^2$. Then the given RBM is the limit of the M/M/1/C models as described in Theorem 4.3. Next note that the M/M/1/C model has the same regenerative structure as RBM, as described on pp. 86–89 of Harrison [1985]. By Theorem 4.3 and the continuous mapping theorem (Theorem 5.1 of Billingsley [1968]), the processes associated with a single cycle converge. In particular, consider the first passage time from 0 to 0 after hitting C applied to $\mathbf{T}_n$ and Z. This mapping, say τ, is a measurable function on D that is continuous almost surely with respect to the underlying Brownian motion X. To see this, consider the first passage times from 0 to C and from C to 0 separately. Focusing only on the upward first passage time (since the downward first passage time is similar), suppose that $Z(0) = X(0) = 0$ and let ξ be the first passage time of Z to C. Then, with probability 1, $X(t)$ exceeds $X(\xi)$ somewhere in every neighborhood to the right of ξ. For any such

sample path, this ensures that, for sufficiently large n, Z_n will first hit C in any specified neighborhood of ξ if Z_n is associated with X_n for which $X_n(t) \to X(t)$ uniformly on bounded intervals, as occurs with the heavy-traffic limit. Hence,

$$[\tau(\mathbf{T}_n), \mathbf{O}_{Jn}(\tau(\mathbf{T}_n)), \mathbf{O}_{Tn}(\tau(\mathbf{T}_n))] \Rightarrow [\tau(Z), L(\tau(Z)), U(\tau(Z))] \tag{52}$$

as $n \to \infty$. Now we want to conclude from (52) that we have convergence of the following moments:

$$E[\tau(\mathbf{T}_n)^k] \to E[\tau(Z)^k] \quad \text{for } k = 1, 2, \tag{53}$$

$$E[\mathbf{O}_{Jn}(\tau(\mathbf{T}_n))^k] \to E[L(\tau(Z))^k] \quad \text{for } k = 1, 2, \tag{54}$$

$$E[\mathbf{O}_{Tn}(\tau(\mathbf{T}_n))^k] \to E[U(\tau(Z))^k] \quad \text{for } k = 1, 2, \tag{55}$$

$$E[\tau(\mathbf{T}_n)\mathbf{O}_{Jn}(\tau(\mathbf{T}_n))] \to E[\tau(Z)L(\tau(Z))], \tag{56}$$

and

$$E[\tau(\mathbf{T}_n)\mathbf{O}_{Tn}(\tau(\mathbf{T}_n))] \to E[\tau(Z)U(\tau(Z))] \quad \text{as } n \to \infty. \tag{57}$$

First, assuming that (53)–(57) do indeed hold, we establish our desired result. We apply the alternative formulas for these quantities in Section 3 and the limits in Theorem 4.5 to obtain the explicit formulas. Let τ_n be the first passage of time of T_n (without normalization) from 0 to 0 after hitting C_n. Then $\tau_n = n\tau(T_n)$,

$$\mathbf{O}_{Jn}(\tau(\mathbf{T}_n)) = \frac{O_{Jn}(\tau_n)}{\sqrt{n}} \quad \text{and} \quad \mathbf{O}_{Tn}(\tau(\mathbf{T}_n)) = \frac{O_{Tn}(\tau_n)}{\sqrt{n}}. \tag{58}$$

Since O_{Jn} is a renewal process,

$$c_{Jn}^2 = \lim_{t \to \infty} \frac{\text{Var } O_{Jn}(t)}{EO_{Jn}(t)}.$$

Then, using regenerative structure (e.g., see Section 3 of Glynn and Whitt [1987]), we obtain

$$c_{Jn}^2 = \frac{\text{Var}\left(O_{Jn}(\tau_n) - \dfrac{\tau_n}{E\tau_n} EO_{Jn}(\tau_n) \right)}{EO_{Jn}(\tau_n)}. \tag{59}$$

Next, introducing the normalization in (44), we obtain

$$\frac{c_{Jn}^2}{\sqrt{n}} = \frac{\text{Var}\left[\mathbf{O}_{Jn}(\tau(\mathbf{T}_n)) - \dfrac{\tau(\mathbf{T}_n)E[\mathbf{O}_{Jn}(\tau(\mathbf{T}_n))]}{E[\tau(\mathbf{T}_n)]}\right]}{E[\mathbf{O}_{Jn}(\tau(\mathbf{T}_n))]}$$

$$\to \frac{\text{Var}\left[L(\tau(Z)) - \dfrac{\tau(Z)E[L(\tau(Z))]}{E[\tau[Z]]}\right]}{E[L(\tau(Z))]} = \alpha^{-1}\sigma_L^2 \quad \text{as } n \to \infty. \tag{60}$$

Now we complete the proof by establishing (53)–(57). First, the convergence (53) follows from p. 151 of Abate and Whitt [1988], where explicit expressions are given for the first two moments. Next, since

$$\frac{EO_{Jn}(\tau_n)}{E\tau_n} = \lambda_n' \quad \text{and} \quad \frac{EO_{Tn}(\tau_n)}{E\tau_n} = r_n',$$

by (50) and (53)

$$\frac{EO_{Jn}(\tau_n)}{\sqrt{n}} \to \alpha E\tau(Z) \quad \text{and} \quad \frac{EO_{Tn}(\tau_n)}{\sqrt{n}} \to \beta E\tau(Z) \quad \text{as } n \to \infty,$$

which implies (54) and (55) for $k = 1$. The rest we establish by uniform integrability; see p. 32 of Billingsley [1968]. Here there are two gaps, because we have not performed two calculations. To establish the uniform integrability, it suffices to show that $E[\mathbf{O}_{Jn}(\tau(\mathbf{T}_n))^3]$ and $E[\mathbf{O}_{Tn}(\tau(\mathbf{T}_n))^3]$ are uniformly bounded in n. Since these two are essentially the same, we only discuss the first. For this purpose, we show that $E[\tau(\mathbf{T}_n)^3]$ and

$$\beta_k\left[\mathbf{O}_{Jn}(\tau(\mathbf{T}_n)) - \frac{\tau(\mathbf{T}_n)E[\mathbf{O}_{Jn}(\tau(\mathbf{T}_n))]}{E[\tau(\mathbf{T}_n)]}\right] \tag{61}$$

are uniformly bounded in n for $k = 2$ and 3, where $\beta_k(Z)$ is the k^{th} cumulant of the random variable Z, i.e., the coefficient of t^k in the power series representation of $\log Ee^{tZ}$; see Section 2 of Whitt [1982a]. For $k = 2$ and 3, the cumulants are the variance and the third central moment. The uniform boundedness of $E[\tau(\mathbf{T}_n)^3]$ in n should follow by differentiating the transform, which is displayed in Theorem 3.4 of Abate and Whitt [1988], just as for (53) with $k = 1$ and 2; this is the first gap. The k^{th} cumulant (61) can be expressed without the normalization as

$$n^{-k/2}\,\beta_k\left[O_{Jn}(\tau_n) - \frac{\tau_n EO_{Jn}(\tau_n)}{E\tau_n}\right]. \tag{62}$$

Since $n\,E\tau_n \to E\tau(Z)$, if we multiply and divide by $E\tau_n$ in (62), then we see that it suffices to show that

$$n^{-(k-2)/2} \frac{\beta_k(O_{Jn}(\tau_n) - \tau_n EO_{Jn}(\tau_n)/E\tau_n)}{E\tau_n} \tag{63}$$

is uniformly bounded in n. However, by regenerative structure, (63) coincides with

$$n^{-(k-2)/2} \lim_{t\to\infty} \frac{\beta_k(O_{Jn}(t))}{t}. \tag{64}$$

For $k = 2$ and 3, the limit in (64) can be expressed in terms of the first three moments of the time between overflows, say J_n as in (2.7) of Whitt (1982a). Using (7) and the reasoning in the proof of Theorem 3.1, we see that $n^{-1/2}EJ_n$, $n^{-3/2}EJ_n^2$, and $n^{-5/2}EJ_n^3$ are uniformly bounded for $k = 1$, 2, and 3, which establishes the desired results; the bound for EJ_n^3 is the second gap. For the case $\rho < 1$, the argument using (7) may be simplified by exploiting the fact that the M/M/1/C busy period is stochastically dominated by the M/M/1/∞ busy period, whose first few moments have simple explicit expressions; see Corollary 3.1.1 of Abate and Whitt (1988). The second and third moments of the busy period are order $(1 - \rho)^{-3}$ and $(1 - \rho)^{-5}$, respectively. ■

Remark. The argument used in the proof of Theorem 4.1 could also be used to establish (50) and (51) for GI/M^X/1/C models, provided that the uniform integrability leading to (53)–(57) can be established.

5. Markov-Chain Approximations

Our fourth approximation scheme for analyzing a token bank is to approximate the general job arrival process, which has been partially characterized by the parameters λ and c_λ^2, by a convenient special renewal process, and then do an exact Markov chain analysis. The first approximating renewal process is a batch-Poisson M^X process having geometric batch sizes. Since this M^X process is a two-parameter renewal process, we choose the two parameters—the batch arrival rate λ^b and the mean batch size m^b—so that the parameters λ and c_λ^2 match; i.e.,

$$\lambda = \lambda^b m^b \quad \text{and} \quad c_\lambda^2 = 2m^b - 1. \tag{65}$$

The batch-size probability mass function is $b(n) = (1 - q)q^{n-1}$, $n \geq 1$, where $m^b = 1/(1 - q)$ or $q = (m^b - 1)/m^b$. Since the batch sizes are always at least 1, $m^b \geq 1$, and this approximation requires $c_\lambda^2 \geq 1$.

After making the M^X approximation, the number of tokens in the dedicated bank coincides exactly with the queue-length process in a GI/M^X/1/C model. The number of tokens in the bank just prior to a token arrival is thus a discrete-time Markov chain (MC), which has been analyzed in the case of deterministic arrivals by Berger[1991a]. From the equilibrium vector of this MC, say $\pi = \{\pi(0), \ldots, \pi(C)\}$, we obtain the exact token blocking probability

$$\frac{r'}{r} = \pi(C). \tag{66}$$

We then invoke the conservation law (2) to obtain the job-blocking probabilities; i.e.,

$$\frac{\lambda'}{\lambda} = 1 - \frac{r - r'}{\lambda}. \tag{67}$$

From the point of view of accuracy with M^X job arrival streams, the exact results (66) and (67) are better than the previous approximations, but (66) is not closed form.

We can use the steady-state distributions at token arrival epochs to compute the steady-state distribution at arbitrary times (and thus at batch arrival points) for the $GI/M^X/1/C$ model. For the special case of Poisson job arrivals (the GI/M/1/C model), there is the relation

$$p(k) = \begin{cases} \rho\pi(k-1), & 1 \le k \le C, \\ 1 - \sum_{j=1}^{C} p(j), & k = 0, \end{cases} \tag{68}$$

where $p(k)$ is the steady-state probability at an arbitrary time; see Heyman and Stidham [1980].

The approximation just presented requires $c_\lambda^2 \ge 1$. If $c_\lambda^2 < 1$, then we could simply act as if $c_\lambda^2 = 1$ and work with the GI/M/1/C model. Alternatively, we could fit a phase-type renewal process and apply a Markov chain analysis to the GI/PH/1/C model, keeping track of the phase of service, as in Berger [1991b]. This is the second class of approximating renewal processes.

Since the token overflow process is a renewal process in the $GI/M^X/1/C$ model, it is natural to use the SCV c_T^2 to further partially characterize the token overflow stream. The interoverflow time is distributed exactly as the first passage time from state C to state C in the ergodic discrete-time MC with states $\{0, 1, \ldots, C\}$ obtained by looking at the $GI/M^X/1/C$ model at (just before) token arrival epochs. The mean interoverflow time is $1/r'$, which we can obtain from (66). From Kemeny and Snell [1959], we obtain

$$c_T^2 = [W]_{C\times C} \cdot (\pi(C))^2 - 1, \tag{69}$$

where $[W]_{C\times C}$ is the $(C \times C)^{\text{th}}$ element of the matrix of the second moments of the first passage times (measured in number of steps of the Markov chain) given by

$$W = M(2Z_{\text{dg}}\,\Delta - I) + 2(ZM - E(ZM)_{\text{dg}}), \tag{70}$$

M is the matrix of mean first passage times and Z is the fundamental matrix, i.e.,

$$M = (I - Z + EZ_{dg})\,\Delta \quad \text{and} \quad Z = (I - P + A)^{-1}, \tag{71}$$

where P is the one-step transition matrix of the Markov chain, A is the square matrix with each row being the equilibrium vector of P, Δ is the diagonal matrix whose diagonal elements are the reciprocal of the equilibrium probabilities, I is the identity matrix, E is the square matrix with all entries equal to 1, and $(\cdot)_{dg}$ is a diagonal matrix whose diagonal equals that of the argument matrix.

The job overflow process associated with a GI/M^X/1/C model is not renewal process, so it is not so easy to analyze or partially characterize. Of course, the mean job interoverflow time is $1/\lambda'$, which we can obtain from (67). We do not give a second parameter for the job overflow stream via this approach.

6. More Approximations for SCVs

We developed approximations for the SCVs of the overflow streams in Sections 3–5. We consider a different approach here, and also consider the stream of admitted jobs. As in Section 4, let $A_J(t)$ and $A_T(t)$ represent, respectively, the number of arriving jobs and tokens in $[0, t]$; and let $O_J(t)$ and $O_T(t)$ represent the number of jobs and tokens to overflow in $[0, t]$. Let c_J^2, c_T^2, and c_A^2 represent the SCVs partially characterizing the job overflow stream, the token overflow stream, and the admitted stream, respectively. Suppose that $A_J(t)$ and $A_T(t)$ are independent renewal processes with rates λ and r, and SCVs c_λ^2 and c_r^2.

If $O_J(t)$ and $O_T(t)$ were renewal processes, then we would have

$$c_J^2 = \lim_{t\to\infty} \frac{\text{Var } O_J(t)}{EO_J(t)} \quad \text{and} \quad c_T^2 = \lim_{t\to\infty} \frac{\text{Var } O_T(t)}{EO_T(t)} \tag{72}$$

(see Whitt [1982a]), so it is natural to use this large-time behavior as a basis for approximations.

First, for λ sufficiently less than r (ρ sufficiently greater than 1), we should have

$$O_T(t) \approx A_T(t) - A_J(t), \tag{73}$$

so that, for suitably large t,

$$c_T^2 \equiv c_T^2(\rho, c_\lambda^2, c_r^2) \approx \frac{\text{Var}(A_T(t) - A_J(t))}{E(A_T(t) - A_J(t))} \approx \frac{\rho c_r^2 + c_\lambda^2}{\rho - 1}. \tag{74}$$

Moreover, in this case the accepted job stream is approximately equal to the job arrival stream, so that $c_A^2 \approx c_\lambda^2$ and the admitted stream is nearly a renewal process if the job arrival stream is.

Second, for r sufficiently less than λ (ρ sufficiently less than 1), we should have

$$O_J(t) \approx A_J(t) - A_T(t), \tag{75}$$

so that, for suitably large t,

$$c_J^2 \equiv c_J^2(\rho, c_\lambda^2, c_r^2) \approx \frac{\text{Var}(A_J(t) - A_T(t))}{E(A_J(t) - A_T(t))} \approx \frac{\lambda c_\lambda^2 + r c_r^2}{\lambda - r} = \frac{c_\lambda^2 + \rho c_r^2}{1 - \rho}. \tag{76}$$

Unlike the case with $\rho > 1$, when $\rho > 1$ the admitted stream is not nearly the same as either arrival stream. For large time, the number of admitted jobs is approximately equal to the number of arrivals taken, so that we should have $c_A^2 \approx c_r^2$ by the asymptotic criterion of (72). However, in shorter time the admitted stream reflects the behavior of the job arrival stream. Thus, if c_λ^2 and c_r^2 are quite different, then we expect that the admitted stream will not nearly be a renewal process.

These approximations for the overflow SCVs provide additional support for the Brownian approximation in Section 4, because both approximations (74) and (76) are consistent with the diffusion approximation in (29): As $\theta C \to -\infty$, $\alpha^{-1}\sigma_L^2$ in (29) is asymptotic to $-2/\theta$, which is (76); as $\theta C \to +\infty$, $\alpha^{-1}\sigma_L^2$ is asymptotic to $2/\theta$, which is (74).

Recall that we already have an algorithm to compute c_T^2 exactly for the GI/M^X/1/C model in Section 5, but (74) is closed form. Note that approximations (74) and (76) apply only to dominant overflow stream (c_J^2 when $\rho < 1$ and c_T^2 when $\rho > 1$). However, these streams are most important in approximations for the multiclass throttle.

7. The Token-Bank Throttle Compared to the Leaky-Bucket Throttle

In this section we compare the token bank to the leaky bucket. First, we define the leaky bucket. With the conventional definition, the leaky bucket has a *drain rate r* and a capacity C. At a job arrival, if the bucket content is less than or equal to $C - 1$, then the job is admitted and the content of the bucket is *incremented* by 1. Otherwise, the job is rejected (or marked and admitted). The content of the bucket drains out at the deterministic rate r. When the bucket is empty, the draining process stops. The draining process starts again upon the next job arrival. The arrival causes the bucket content to be incremented by 1, and a new busy period of the bucket begins. Thus, the time epochs at which a unit of content drains out do not remain synchronous throughout time, but rather undergo a phase shift each time the bucket empties.

In contrast, for the token-bank throttle, the token arrival process continues to run independent of the state of the bank and, in particular, when the bank is full. Thus, the token arrival epochs do remain synchronous throughout time.

To compare the leaky bucket to the token bank, it is convenient to introduce a second definition of the leaky bucket that is isomorphic to the classic one above. The new definition is expressed in terms of a modified token bank. In particular, the leaky-bucket throttle is isomorphic to a token bank that, as usual, decrements at job admissions and increments at token arrivals, but where the deterministic, evenly spaced token arrival process stops when the bank is full and starts again at the next job arrival. Moreover, to analyze the

leaky bucket, it is useful to assume that during intervals when the bank is full, tokens arrive according to a Poisson process with rate r. All of these "Poisson-arriving" tokens are lost (overflow) and do not affect the state of the bank. By a regenerative argument, it is easy to see that the overall token arrival rate is exactly r with this modification. Below we analyze the leaky bucket as this modified token bank.

Budka [1990] established a very general comparison result, showing that, for given capacity, given constant deterministic token arrival and rate and given general job arrival process, that the token-bank throttle admits at least as many jobs as the leaky-bucket throttle. We complement Budka's result in two ways. First, we consider more general token arrival processes. (Recall that general token arrival processes appear in the multiclass throttle. Renewal token arrival processes are a natural generalization in the modified-token-bank formulation of the leaky bucket. This generalization corresponds in the leaky bucket to a non-constant drain rate where the times to deplete successive units is a renewal process.) Second, we show that although the standard D/M/1/C token bank blocks fewer jobs then the leaky bucket for a given capacity, the reduction in blocking is not much when the capacity is large. We quantify this qualification, by showing that the leaky bucket blocks fewer jobs than the token bank if the capacity in the token bank is reduced by 1.

Let subscripts LB and TB designate the leaky bucket and the token bank, respectively. As in Section 5, we use GI/M^X/1/C to refer to a renewal (GI) token arrival stream, a batch Poisson M^X job arrival stream, and capacity C. For the GI/M^X/1/C model of the token bank, we focus on the embedded discrete-time MC at (just before) token arrivals, with stationary probability mass function π. Let $\tau(j, k)$) be the mean first passage time to state k from state j in this MC. As in Section 5, let λ^b be the batch arrival rate and let $\rho_b = r/\lambda^b$.

In this section we state some results for the token-blocking probability r'/r, but corresponding results for the job-blocking probabilities λ'/λ follow from the conservation law (2), since both models have the same λ and r.

THEOREM 7.1. For the GI/M^X/1/C model, the token-blocking probability is

$$\frac{r'_{\mathrm{TB}}}{r} = \pi(C) = \frac{1}{\tau(C-1, C)}, \tag{77}$$

whereas in the leaky bucket it is

$$\frac{r'_{\mathrm{LB}}}{r} = \begin{cases} \dfrac{\rho_b}{\rho_b + \sum_{k=1}^{C-2} \tau(C-1-k, C-1)b(k) + \tau(0, C) \sum_{k=C-1}^{\infty} b(k)} & \text{if } C \geq 2, \\ \dfrac{\rho_b}{1+\rho_b} & \text{if } C = 1 \end{cases} \tag{78}$$

where $b(k)$ is the probability that an arriving job batch is of size k.

Proof. The formula for the token bank has been given in (66). It is well known that $\pi(C) = 1/\tau(C, C)$ and $\tau(C-1,C)=\tau(C, C)$ because the starting state after the initial token

arrival is C in both cases. For the leaky bucket, we do a regenerative analysis. Let the regeneration epochs occur upon a job arrival to a full token bank. By convention, the GI token arrival stream is turned on at this point. Since the arriving job takes a token away, the initial state after the job arrival is $C - k$ with probability $b(k)$, which is the same as if a token were to arrive and find $C - k - 1$ tokens in the bank (assuming that $C \geq k + 1$). Hence, the expected number of tokens to arrive before the bank is again full is

$$\sum_{k=1}^{C-2} \tau(C - 1 - k, C - 1)b(k) + \tau(0, C) \sum_{k=C-1}^{\infty} b(k) \quad \text{if } C \geq 2.$$

After the final token arrival, the bank will be full. At this point, the GI token arrival stream is turned off and the Poisson token arrival stream is turned on. The expected time until the next job arrival is $(\lambda^b)^{-1}$, where λ^b is the batch arrival rate in (65). Using the special model for the leaky bucket above, we see that the expected number of rejected tokens in this period is $\rho_b = r/\lambda^b$. Thus, the long run proportion of rejected tokens is as given in (78). ■

We now make comparisons between the token bank and the leaky bucket. For this purpose, recall that one cdf G_1 is less than or equal to another G_2 in the *convex stochastic ordering*, denoted by $G_1 \leq_c G_2$, if

$$\int f(x)\, dG_1(x) \leq \int f(x)\, dG_2(x) \tag{79}$$

for all convex real-valued functions f for which the expectations are well defined; see Ross [1982]. Since $f(x) = x$ and $f(x) = -x$ are both convex, convex stochastic order implies equal means. Another ordering is *Laplace transform stochastic ordering*, denoted by $G_1 \leq_L G_2$, which occurs if (79) holds for all f of the form $f(x) = e^{-\lambda x}$ for $\lambda > 0$; e.g., see Whitt [1984a]. Obviously $G_1 \leq_L G_2$ if $G_1 \leq_c G_2$.

The following is an explicit comparison for C=1. Note that the comparison strongly depends on the token interarrival time distribution. In particular, Budka's [1990] comparison result does not extend to arbitrary token processes.

THEOREM 7.2. For the GI/M^X/1/C model with $C = 1$,

$$\frac{r'_{TB}}{r} = \pi(1) = \frac{1}{\tau(0, 1)} = \int_0^\infty e^{-\lambda^b x}\, dG(x), \tag{80}$$

where G is the cdf of the token interarrival time. Hence, $\lambda'_{TB}(G_1) \leq \lambda'_{TB}(G_2)$ if $G_1 \leq_L G_2$. Moreover,

$$\begin{aligned} 1 - \rho + \rho e^{-\rho_b^{-1}} &= \frac{\lambda'_{TB}(D)}{\lambda} \leq \frac{\lambda'_{TB}(G)}{\lambda} \leq \frac{\lambda'_{TB}(M)}{\lambda} \\ &= \frac{\lambda'_{LB}(G)}{\lambda} = 1 - \rho + \frac{\rho\rho_b}{1 + \rho_b} \quad \text{if } G \leq_c M, \end{aligned} \tag{81}$$

while

$$\frac{\lambda'_{TB}(G)}{\lambda} \geq \frac{\lambda'_{TB}(M)}{\lambda} = \frac{\lambda'_{LB}(G)}{\lambda} = 1 - \rho + \frac{\rho\rho_b}{1+\rho_b} \quad \text{if } G \geq_c M. \tag{82}$$

Proof. First, note that for both the token bank and the leaky bucket, the conservation law (2) implies that

$$\frac{\lambda'}{\lambda} = 1 - \rho + \rho\frac{r'}{r}.$$

For the leaky bucket, r'_{LB}/r is given by (78). Thus

$$\frac{\lambda'_{LB}(G)}{\lambda} = 1 - \rho + \frac{\rho\rho_b}{1+\rho_b}.$$

For the token bank, note that

$$\tau(0, 1) = p + (1-p)(1 + \tau(0, 1)),$$

where p is the probability of no job arrivals in a token interarrival time; i.e.,

$$p = \int_0^\infty e^{-\lambda^b x}\, dG(x). \tag{83}$$

Hence, $p\tau(0, 1) = 1$ and we have established (80). From (80), we immediately have $r'_{TB}(G_1) \leq r'_{TB}(G_2)$ when $G_1 \leq_L G_2$. For the M/M^X/1/C model, p in (83) is $\rho_b/(1+\rho_b)$ and thus $\lambda'_{TB}(M) = \lambda'_{LB}(G)$. For the D/M^X/1/C model, p is $e^{-\rho_b^{-1}}$. Finally, note that $G_1 \leq_c G_2$ if G_1 is the cdf of a deterministic distribution with a unit point mass at the mean of G_2. ■

Theorem 7.2 suggests that the token bank produces less blocking than the leaky bucket when the token stream is smooth. Our next result establishes an ordering the other way independent of the token interarrival-time distribution, when the capacity of the token bank is decreased by 1. We also prove that the fluid approximation is a lower bound to the token-blocking probability when $\rho > 1$. Let r'_F be the fluid approximation in (1).

THEOREM 7.3. (a) For the GI/M^X/1/C model with $\rho_b < 1$ and $C \geq 2$, $\lambda'_{LB}(C) \leq \lambda'_{TB}(C-1)$.

(b) For the GI/M/1/C model with $C \geq 2$, $\lambda'_{LB}(C) \leq \lambda'_{TB}(C-1)$.

(c) For the GI/M/1/C model with $\rho > 1$,

$$\frac{r'_{TB}(C)}{r} \geq \lim_{C\to\infty} \frac{r'_{TB}(C)}{r} = \frac{\rho - 1}{\rho} = \frac{r'_F}{r}.$$

Proof. (a) First note that $\tau(0, C) \geq \tau(C-2, C-1)$ and $\tau(C-k-1, C-1) \geq \tau(C-2, C-1)$ for all $k \geq 1$, so that from (78)

$$\frac{r'_{\mathrm{LB}}(C)}{r} \leq \frac{\rho_b}{\rho_b + \tau(C-2, C-1)} \quad \text{if } C \geq 2. \tag{84}$$

Note that $\tau(C-2, C-1)$ appears on both (77) and (84) when we consider $r'_{\mathrm{LB}}(C)$ and $r'_{\mathrm{TB}}(C-1)$. Hence, $r'_{\mathrm{LB}}(C) \leq r'_{\mathrm{TB}}(C-1)$ whenever

$$(\rho_b - 1)\tau(C-2, C-1) \leq \rho_b. \tag{85}$$

Inequality (85) is obviously satisfied when $\rho_b \leq 1$. (b) For the GI/M/1/C model, the argument above covers the case $\rho \leq 1$. When $\rho > 1$, we use the fact that $\tau(C-2, C-1)$ is increasing in C, which is easily shown by a coupling argument. (In particular, to compare $\tau(k, k+1)$ and $\tau(k-1, k)$, use the same arrival times and service times. If the sample path starting in k never reaches 0, then $\tau(k, k+1) = \tau(k-1, k)$ for that sample path. However, if the sample path starting in k hits 0, then the two paths move together afterwards, so that they both reach k at the same time and $\tau(k-1, k) < \tau(k, k+1)$ for that sample path.) Moreover, the limit as $C \to \infty$ is equivalent to an M/G/1 queue. In particular, by considering C minus the number of tokens in the bank, we see that $\tau(C-1, C) \to \tau(\infty)$ as $C \to \infty$ when $\rho > 1$, where $\tau(\infty)$ is the mean number of customers served in a busy period of an M/G/1 queue with arrival rate λ, mean service time r^{-1} and traffic intensity ρ^{-1}. Hence, $\tau(\infty) = (1-\rho^{-1})^{-1} = \rho/(\rho-1)$ and

$$\tau(C-2, C-1) \leq \lim_{C\to\infty} \tau(C-1, C) = \frac{\rho}{\rho-1}, \tag{86}$$

which equals $\rho_b/(\rho_b - 1)$ in this case. Finally, (86) establishes (c). ■

We now show that the token bank blocks fewer jobs than the leaky bucket for common parameters in the D/M/1/C model, which is a very special case of Budka's [1990] result. We *conjecture* that this result remains true for the GI/M^X/1/C model provided the token interarrival-time cdf G satisfies $G \leq_c M$. The necessity of some condition is shown by Theorem 7.2.

THEOREM 7.4. For the D/M/1/C model, $\lambda'_{\mathrm{TB}}(C) \leq \lambda'_{\mathrm{LB}}(C)$.

Proof. By (77) and (78), it suffices to show that

$$\tau(C-1, C) - 1 - \rho^{-1}\tau(C-2, C-1) \geq 0. \tag{87}$$

By considering the possible Poisson events in the first token interarrival time, we can write

$$\begin{aligned}\tau(C-1, C) \geq\ & e^{-\rho^{-1}} + \rho^{-1} e^{-\rho^{-1}}(1 + \tau(C-1, C)) \\ & + (1 - e^{-\rho^{-1}} - \rho^{-1} e^{-\rho^{-1}})(1 + \tau(C-2, C))\end{aligned}$$

$$\geq e^{-\rho^{-1}} + \rho^{-1} e^{-\rho^{-1}} (1 + \tau(C - 1, C)) + (1 - e^{-\rho^{-1}} - \rho^{-1} e^{-\rho^{-1}})(1 + \tau(C - 2, C - 1) + \tau(C - 1, C)).$$

Hence,

$$e^{-\rho^{-1}} \tau(C - 1, C) \geq 1 + (1 - e^{-\rho^{-1}} - \rho^{-1} e^{-\rho^{-1}}) \tau(C - 2, C - 1),$$

so that

$$\tau(C - 1, C) \geq e^{\rho^{-1}} + (e^{\rho^{-1}} - 1 - \rho^{-1}) \tau(C - 2, C - 1).$$

However, by doing a sample path comparison it is possible to show that $\tau(j + m, k + m)$ for $j < k$ is increasing in m, so that

$$\tau(C - 2, C - 1) \geq \tau(0, 1) = e^{\rho^{-1}}.$$

Therefore,

$$\tau(C - 1, C) - 1 - \rho^{-1} \tau(C - 2, C - 1) \geq (e^{\rho^{-1}} - 1) + (e^{2\rho^{-1}} - e^{\rho^{-1}} - \rho^{-1} e^{\rho^{-1}}) - \rho^{-1} e^{\rho^{-1}} \geq f(\rho^{-1}) \equiv e^{2\rho^{-1}} - 2\rho^{-1} e^{\rho^{-1}} - 1 \geq 0$$

for all $\rho \geq 0$

because $f(0) = 0$ and

$$f'(x) = 2e^x (e^x - 1 - x) > 0 \qquad \text{for all } x > 0. \quad \blacksquare$$

We conclude this section by giving some numerical results. Table 1 compares the job blocking probabilities in the token bank and the leaky bucket for the D/M/1/C model with capacities 5 and 11. The (exact) leaky bucket results are obtained by applying results for the dual M/D/1/C queue with traffic intensity ρ^{-1} in Kühn [1976] while the (exact) token bank results are obtained by the MC analysis in Section 5.

To provide a better understanding, the results are related to the fluid approximation. First the fluid approximation (1) itself is displayed, it yields $\lambda' = 0$ when $\lambda < r$ and $\lambda' > 0$ when $\lambda > r$. Second, the other results are displayed after subtracting fluid approximation. Thus, the other values describe the *excess blocking* over the fluid approximation.

From Table 1, we see that the fluid approximation provides much of the story. The relative error in throughput $\lambda - \lambda'$ for the fluid approximation certainly is not great; the worst case occurs when $\rho = 1$. The excess blocking probabilities for the token bank and the leaky bucket are similar, but not too close. For example, with capacity $C = 5$, using the exact value of one as an approximation for the other would yield relative errors in the

Table 1. A comparison of job-blocking probabilities λ'/λ in the token bank and the leaky bucket for the case of a deterministic token at rate r and Poisson job arrivals at rate λ. The desired fluid rate is subtracted in each case. The cases $C = 10$ and 11 are both incuded to illustrate Theorem 7.3 and 7.4.

		Bank capacity				
		$C = 11$		$C = 10$	$C = 5$	
λ/r	Fluid	Token bank − fluid	Leaky bucket − fluid	Token bank − fluid	Token bank − fluid	Leaky bucket − fluid
0.5	0				0.00124	0.00218
0.6	0				0.00429	0.00667
0.7	0	0.00020	0.00028	0.00040	0.0120	0.01671
0.8	0	0.00191	0.00235	0.00295	0.0279	0.0353
0.9	0	0.0117	0.0132	0.0148	0.0550	0.0644
1.0	0	0.0441	0.0462	0.0484	0.0938	0.1035
1.2	0.16667	0.00237	0.00289	0.00349	0.0259	0.0325
1.4	0.28571	0.00009	0.00013	0.00018	0.00644	0.00951
1.6	0.37500				0.00158	0.00276

excess blocking probability of 10%–50% over the range for $0.7 \leq \rho \leq 1.4$. Finally, the fact that $\lambda'_{TB}(C) \leq \lambda'_{LB}(C) \leq \lambda'_{TB}(C-1)$ is illustrated for the case $C = 11$.

8. The Accuracy of the Approximations

In this section we evaluate the approximations by making numerical comparisons, giving special attention to the Brownian approximation. We primarily focus on the steady-state mean number of tokens in the token bank, the job and tokens overflow rates, and the SCVs approximately characterizing the overflow processes.

We use several sources of numerical values. First, we solve $D/M^X/1/C$ and $D/H_2/1/C$ models exactly using Markov chain analysis, as described in Section 5. Of course, the standard D/M/1/C model of a dedicated token bank is a special case of both, so the two analyses provide checks on each other. Next, we use the exact formulas for the M/M/1/C model in Section 3. We also examine the direct Brownian approximation and two refinements: the M/M/1/C refinement in Section 4.3 and the GI/G/1/∞ refinement in Section 4.4. For additional comparisons, we use tabled results for GI/G/1/C queues in Kühn [1976] and Seelen, Tijms, and van Hoorn [1985]. Finally, we use a special-purpose FORTRAN simulation of the token bank, primarily to evaluate the variability of nonrenewal overflow processes.

8.1 The Steady-State Mean

To understand the numerical accuracy of the direct Brownian approximation, it is natural to start by considering the M/M/1/∞ model with $\rho < 1$. For this special case, there is the simple relation between the approximate mean and the exact value:

$$EZ(\infty) = \frac{-1}{\theta} = \frac{1+\rho}{2(1-\rho)} = \frac{\rho}{1-\rho} + \frac{1}{2}. \tag{88}$$

Hence, the direct Brownian approximation for the M/M/1/∞ mean always has an error of exactly 1/2. This suggests that we should expect the direct Brownian approximation to consistently achieve accuracy within about 0.5 in the steady-state mean when the SCVs c_λ^2 and c_r^2 are not too different from 1. Numerical evidence shows that this is the case for the M/M/1/C model for all ρ and C.

Formula (88) also indicates that the direct Brownian approximation for the mean can benefit from some refinement. Of course, any error for the mean in the M/M/1/C model is eliminated completely by the M/M/1/C and GI/G/1/∞ refinements in Sections 4.3 and 4.4. Another approach to the discrepancy in (88) is to use a different diffusion coefficient; e.g., instead of (20) let $\sigma^2 = rc_r^2 + \rho\lambda c_\lambda^2 = r(c_r^2 + c_\lambda^2)$. This refinement eliminates the error in (88), but it does not work so well for the overflow processes; see Sections 8.2 and 8.5 for further discussion.

We now investigate how well the Brownian approximations work for SCVs different from 1. A relatively nice case is the $E_2/E_2/1/C$ queue, for which both the interarrival times and service times have E_2 distributions (Erlang of order 2, convolution of two exponential distributions), so that $c_r^2 = c_\lambda^2 = 0.50$. Table 2 compares the approximations for the mean with exact values from Seelen, Tijms and van Hoorn [1985] for the case of $C = 11$. For this case the direct approximation is quite good. Indeed, the direct approximation is slightly better than the M/M/1/C refinement, and only slightly worse than the GI/G/1 refinement.

Table 3 compares several expressions for the steady-state mean number of tokens in the token bank for several values of the traffic intensity ρ when the capacity C is 100 and the SCVs are $c_\lambda^2 = 10$ and $c_r^2 = 0$. This is a much more stressful test than the $E_2/E_2/1/11$ queue from the point of view of the SCVs, but it is easier from the point of view of the capacity.

Table 2. A comparison of approximations with the exact value of the steady-state mean for the $E_2/E_2/1/C$ queue (not token bank) with $C = 11$. The exact values come from Seelen, Tijms, and van Hoorn [1985].

		Brownian approximations			
ρ	$E_2/E_2/1/C$ exact	Direct	M/M/1/C refinement	GI/G/1/∞ refinement	M/M/1/C
0.5	0.70	0.75	0.50	0.71	1.00
0.6	0.98	1.00	0.76	0.99	1.47
0.7	1.42	1.42	1.21	1.41	2.16
0.8	2.21	2.17	2.05	2.14	3.11
0.9	3.63	3.55	3.50	3.18	4.28
1.0	5.62	5.50	5.50	5.50	5.50
1.1	7.43	7.29	7.40	7.36	6.61
1.2	8.61	8.46	8.67	8.49	7.52
1.3	9.28	9.09	9.32	9.13	8.20
1.5	9.91	9.75	10.05	9.75	9.09
2.0	10.43	10.25	10.50	10.29	10.00
Average absolute error		0.10	0.13	0.08	0.68

Table 4. A comparison of the expressions for the mean number of tokens in the token bank in the case of capacity $C = 10$ and $c_r^2 = 0$. Two job arrival SCVs are considered: $c_\lambda^2 = 1$ and 4. The $D/H_2^b/1/C$ values are treated as exact for the average absolute error analsyis.

				Brownian approximations			
$\rho = r/\lambda$	c_λ^2	$D/M^X/1/C$	$D/H_2^b/1/C$	Direct	M/M/1/C refinement	GI/G/1 ∞ refinement	M/M/1/C
0.50	1	0.255	0.255	1.00	0.67	0.625	0.99
	4	1.54	1.53	3.11	2.08	1.33	0.99
0.60	1	0.48	0.48	1.25	0.94	0.87	1.46
	4	2.30	2.11	3.43	2.60	2.06	1.46
0.70	1	0.87	0.87	1.64	1.37	1.26	2.11
	4	3.14	2.73	3.79	3.17	2.89	2.11
0.80	1	1.61	1.61	2.32	2.09	1.96	2.97
	4	3.98	3.38	4.18	3.77	3.67	2.97
0.90	1	2.92	2.92	3.44	3.29	3.24	3.97
	4	4.75	4.05	4.58	4.40	4.37	3.97
1.00	1	4.70	4.70	5.00	5.00	5.00	5.00
	4	5.41	4.72	5.00	5.00	5.00	5.00
1/0.90	1	6.43	6.43	6.72	6.85	6.68	6.03
	4	6.02	5.45	5.46	5.64	5.65	6.03
1/0.80	1	7.72	7.72	8.07	8.76	7.86	7.03
	4	6.63	6.31	6.02	6.40	6.44	7.03
1/0.70	1	8.49	8.49	8.84	9.03	8.55	7.89
	4	7.22	7.25	6.67	7.72	7.31	7.89
1/0.60	1	8.95	8.95	9.25	9.43	8.97	8.54
	4	7.77	8.16	7.37	8.01	8.16	8.54
1/0.50	1	9.25	9.25	9.50	9.67	9.25	9.01
	4	8.26	8.85	8.07	8.70	8.82	9.01
Average absolute error							
	1	0.00		0.49	0.45	0.23	0.73
	4	0.41		0.73	0.28	0.16	0.46

Table 4 displays corresponding mean values for the case of capacity $C = 10$, $c_r^2 = 0$, and job SCVs $c_\lambda^2 = 4$ and 1. For $c_\lambda^2 = 1$, the $D/M^X/1/C$ and $D/H_2^b/1/C$ models both reduce to the D/M/1/C model, so that the values agree. By comparing the cases $c_\lambda^2 = 4$ and $c_\lambda^2 = 1$, we see the impact of the job arrival SCV c_λ^2. From the heavy-traffic limit theorem, we would not expect the Brownian approximation to perform as well when the capacity is reduced from 100 to 10, and this is the case if we use a criterion of relative error. Again the refinements help.

It is important to note that the $D/H_2^b/1/C$ and $D/M^X/1/C$ means are for the embedded Markov chain obtained by looking at the bank just prior to each token arrival. In contrast, we regard the diffusion approximation as being most appropriate to describe the distribution at an arbitrary time in equilibrium. (However, this assertion is based on experience

Table 3. A comparison of expressions for the mean number of tokens in the token bank in the case of capacity $C = 100$ with SCVs $c_\lambda^2 = 10$ and $c_r^2 = 0$. The $D/M^X/1/C$ and $D/H_2^b/1/C$ values are the exact values just prior to token arrivals. The $D/H_2^b/1/C$ model is treated as exact for the average absolute error analysis.

			Browning approximations			
$\rho = r/\lambda$	$D/M^X/1/C$	$D/H_2^b/1C$	Direct	M/M/1/C refinement	GI/G/1/∞ refinement	M/M/1/C
0.50	4.5	6.7	10.0	6.7	2.8	1.0
0.60	7.0	7.2	12.5	9.4	4.9	1.5
0.70	11.2	11.2	16.4	13.5	8.6	2.3
0.80	19.0	18.5	23.1	20.6	16.3	4.0
0.90	33.2	31.2	34.3	32.5	31.5	9.0
0.95	42.4	39.9	41.8	40.8	40.9	18.4
0.99	50.0	47.6	48.3	48.1	48.2	41.6
1.00	51.9	49.6	50.0	50.0	50.0	50.0
1/0.99	53.7	51.6	51.7	51.9	51.8	58.4
1/0.95	60.9	59.7	58.6	59.7	59.5	81.6
1/0.90	69.1	69.6	67.2	68.9	69.8	91.0
1/0.80	81.2	84.6	80.7	82.8	85.9	96.0
1/0.70	88.2	92.5	88.3	90.4	93.1	97.7
1/0.60	92.7	96.4	92.5	94.4	96.3	98.5
1/0.50	94.8	98.3	95.0	96.7	97.9	99.0
Average absolute error						
	2.1		2.9	1.2	1.1	10.3

The first value in Table 3 is the $D/M^X/1/C$ value, which is based on a geometric batch-size distribution, so that the job arrival process is a renewal process with $c_\lambda^2 = 10$. Next comes the $D/H_2^b/1/C$ value, which is based on a hyperexponential distribution (mixture of two exponential distributions) with balanced means and SCV $c_\lambda^2 = 10$; i.e., the cdf is

$$F(t) = 1 - p_1 e^{-t/m_1} - p_2 e^{-t/m_2} \tag{89}$$

where $p_1 + p_2 = 1$ and $p_1 m_1 = p_2 m_2$.

Since the $D/M^X/1/C$ and $D/H_2^b/1/C$ models are both D/GI/1/C models with $c_\lambda^2 = 10$, they both can be considered as the "exact" values in this case. As discussed in Whitt [1984a,b] and Klincewicz and Whitt [1984], when there is more than one possible exact value, it is desirable to compare the approximation to the set of possible exact values; the two cases here give a rough idea of the set. The fact that these values are so close when $C = 100$ and ρ is near 1 is evidence of the invariance principle associated with heavy-traffic limit; i.e., the heavy-traffic limit depends on the two probability distributions in the GI/GI/1/C model only through their first two moments. However, there is quite a range for the two exact values for ρ away from 1.

Also displayed in Table 3 are the direct Brownian approximation, the two refined Brownian approximations and the M/M/1/C formula based on the given ρ. Since $c_\lambda^2 = 10$, we should not expect that the M/M/1/C model would provide a good approximation, and it does not, as illustrated by the large error at $\rho = 0.90$. The direct Brownian approximation performs reasonably well, given the range of exact answers. Indeed, *the direct Brownian approximation is about as good an approximation for the* $D/H_2^b/1/C$ and $D/M^X/1/C$ *models as each is for the other.* The refinements help, but not uniformly so. For example, the GI/G/1/∞ refinement does not do very well for small ρ.

and intuition. The Brownian approximations obtained via heavy-traffic limits are actually the same for the content at arrival epochs and the content at arrival times.) By PASTA, see Wolff [1982], the distribution of the state seen by arrivals is the same as the distribution of the state at arbitrary times in the case of Poisson arrivals. Moreover, in the heavy-traffic limit the two distributions are identical. However, the mean at an arbitrary time in the $D/M^X/1/C$ and $D/H_2^b/1/C$ models, will necessarily be somewhat larger than the mean of the embedded chain (by an amount less than 1).

For the case $c_\lambda^2 = 1$, the $D/H_2^b/1/C$ and $D/M^X/1/C$ models both reduce to the D/M/1/C model, for which we can easily relate the distribution at an arbitrary time to the distribution at an arrival epoch, using (68). In particular, if m_a is the mean of the distribution π obtained by considering the embedded chain at arrival epochs, while m_t is the mean at an arbitrary time, then from (68) we obtain

$$m_t = \rho[m_a + 1 - (C + 1)\pi(C)] \tag{90}$$

and

$$m_a = \rho^{-1} m_t - 1 + (C + 1)\frac{r'}{r}. \tag{91}$$

If we apply (91) to the diffusion approximation, then the accuracy is better. For example, with the GI/G/1/∞ refinement to the case $c_\lambda^2 = 1$ in Table 4, we obtain approximate mean values of 0.25 when $\rho = 0.5$, 0.45 when $\rho = 0.6$, 0.80 when $\rho = 0.7$, 1.45 when $\rho = 0.8$ and 2.50 when $\rho = 0.9$. The average absolute error over these five cases has been reduced from 0.36 to 0.14.

Our analysis of the mean indicates that the direct Brownian approximation provides a reasonable rough approximation, capturing the effect of the parameters used, but for best numerical accuracy we would use the GI/G/1/∞ refinement.

We also calculated the steady-state distribution of the number of tokens in the bank at token arrival epochs for the examples in Tables 2 and 3. The diffusion approximation suggests that the ratios of successive probability mass function values, $p(n + 1)/p(n)$, should be nearly constant. We found this to be the case except at the end points, where there was a significant change.

8.2 *Overflow Rates*

Since $-\theta^{-1}$ is the direct RBM approximation for the steady-state mean in the G/G/1/∞ model when $\rho < 1$, and since θ appears in the exponent of the RBM overflow-rate formulas (25) and (26), we anticipate that the approximation for the *exponent* in the overflow-rate formulas has accuracy comparable to the accuracy of the mean. That is, the relative error in the *logarithm* of the blocking probability should be similar to the relative error in the mean, and this is shown to be the case in the examples.

As we did for the mean, we start by considering the direct Brownian approximation for the M/M/1/C model. As we should expect from Section 4, the direct Brownian approximations

for the overflow rates perform quite well under heavy-traffic conditions, improving as C increases and $|1 - \rho|$ decreases with $|(1 - \rho)C|$ fixed, but the performance of the direct Brownian approximation for the overflow rates deteriorates when $|(1 - \rho)C|$ is large. The direct Brownian approximation performs spectacularly well when $C = 1000$, $\rho = 0.999$, and $(1 - \rho)\ C = 1$ (exact: $r' = 0.0005806$; approximation: $r' = 0.0005815$) and when $C = 100$, $\rho = 0.99$, and $(1 - \rho)\ C = 1$ (exact: $r' = 0.00568$; approximation: $r' = 0.00577$). The direct Brownian approximation perform pretty well when $C = 100$, $\rho = 0.9$, and $(1 - \rho)\ C = 10$ (exact: $r' = 0.24 \times 10^{-5}$; approximation: $r' = 0.27 \times 10^{-5}$), but less well when $C = 10$, $\rho = 0.5$, and $(1 - \rho)\ C = 5$ (exact $r' = 0.24 \times 10^{-3}$; approximation: $r' = 0.64 \times 10^{-3}$), and when C=100, $\rho = 0.5$, and $(1 - \rho)\ C = 50$ (exact $r' = 0.20 \times 10^{-30}$; approximation: r'=0.56 x 10^{-29}). Thus we anticipate that the accuracy of the Brownian approximation for the blocking probabilities will decrease as $|(1 - \rho)C|$ increases for fixed ρ or C. More generally, we anticipate that the accuracy will decrease as $|(1 - \rho)C|$ increases, but the third and fourth examples above show that this does not always hold.

We remark that the refined approximation based on $\sigma^2 = r(c_r^2 + c_\lambda^2)$, which yields $-\theta^{-1} = \rho/(1 - \rho)$ for M/M/1/∞, does not work as well for the overflow rates. For example, when $C = 100$, $\rho = 0.5$, and $(1 - \rho)C = 50$, this "refined" approximation is 0.2 $\times 10^{-45}$, which is in error by a factor of 10^{15}.

Tables 5–7 compare several expressions for the overflow rates. Paralleling Tables 2–4, Table 5 is for the $E_2/E_2/1/C$ queue with $C = 11$; Table 6 is for the case $C = 100$, $c_r^2 = 0$, and $c_\lambda^2 = 10$; and Table 7 is for the case C=10, $c_r^2 = 0$, and $c_\lambda^2 = 4$ and 1. Recall that Table 1 displays blocking probabilities for the token-bank throttle and the leaky-bucket throttle, from which one can calculate overflow rates. Given that we would treat these two throttles the same with Brownian approximation, Table 1 indicates limitations on the accuracy we can hope to achieve.

In Tables 5–7, we only display values for the smaller overflow rate (r' when $r < \lambda$ and λ' when $\lambda < r$); the other overflow rate can be obtained from the conservation law (2). Using a relative error criterion, the larger overflow rates are much easier to approximate accurately because of the deterministic exact component. For example, suppose $r < \lambda$, so that we estimate r'. By the conservation law (2), $\lambda' = (\lambda - r) + r'$. Let subscripts *ex* and *ap* designate the exact and approximate values. Then the relative error in λ' is

Table 5. A comparison of approximations for the overflow rates with exact values for the $E_2/E_2/1/C$ queue (not token bank) with $C = 11$ and $\lambda = 1$ from Seelen, Tijms, and van Hoorn [1985].

			Brownian approximations			
r	Ovfl. rate	$E_2/E_2/1/C$ exact	Direct	M/M/1/C refinement	GI/G/1/∞ refinement	M/M/1/C
0.6		0.000006	0.0000067	0.000004	0.000023	0.00087
0.7	r'	0.000112	0.000127	0.000085	0.00025	0.00421
0.8		0.00144	0.00152	0.00118	0.00196	0.0148
0.9		0.0119	0.0110	0.0095	0.0110	0.0394
1.0		0.0453	0.0455	0.0417	0.0417	0.0833
1.1		0.0139	0.0140	0.0121	0.0130	0.0468
1.2	λ'	0.00365	0.00373	0.00301	0.00436	0.0253
1.3		0.00097	0.00097	0.00073	0.00139	0.0135
1.5		0.000075	0.000075	0.000047	0.000175	0.00389

Table 6. A comparison of expressions for the overflow rates in the case of capacity $C = 100$ with SCVs $c_\lambda^2 = 10$ and $c_r^2 = 0$. The job overflow rate λ' is displayed when $\lambda \leq r$, while the token overflow rate r' is displayed when $r < \lambda$.

					Brownian approximations			
r	λ	Ovfl. rate	D/M^X/1/C	D/H$_2^b$/1/C	Direct	M/M/1/C refinement	GI/G/1/∞ refinement	M/M/1/C
0.50	1.00		0.55×10^{-9}	0.31×10^{-6}	0.23×10^{-4}	0.80×10^{-6}	0.27×10^{-2}	0.20×10^{-30}
0.60	1.00		0.40×10^{-6}	0.13×10^{-4}	0.13×10^{-3}	0.27×10^{-4}	0.20×10^{-2}	0.26×10^{-22}
0.70	1.00		0.40×10^{-4}	0.23×10^{-4}	0.75×10^{-3}	0.36×10^{-4}	0.26×10^{-2}	0.7×10^{-16}
0.80	1.00	r'	0.00109	0.00212	0.00373	0.00272	0.0056	0.3×10^{-10}
0.90	1.00		0.0108	0.0122	0.0157	0.0140	0.0154	0.24×10^{-5}
0.95	1.00		0.0248	0.0249	0.0291	0.0276	0.0271	0.00028
0.99	1.00		0.0423	0.0408	0.0452	0.0445	0.0439	0.0057
1.00	1.00		0.0476	0.0456	0.0500	0.0495	0.0495	0.010
1.00	0.99		0.0428	0.0403	0.0447	0.0440	0.0435	0.0057
1.00	0.95		0.0269	0.0230	0.0268	0.0254	0.0253	0.00028
1.00	0.90	λ'	0.0138	0.0094	0.0122	0.0108	0.0123	0.24×10^{-5}
1.00	0.80		0.0030	0.00065	0.00136	0.0010	0.0020	0.3×10^{-10}
1.00	0.70		0.52×10^{-3}	0.66×10^{-5}	0.57×10^{-4}	0.27×10^{-4}	0.15×10^{-3}	0.7×10^{-16}
1.00	0.60		0.80×10^{-4}	0.10×10^{-8}	0.65×10^{-6}	0.13×10^{-6}	0.34×10^{-5}	0.26×10^{-22}
1.00	0.50		0.11×10^{-4}	0.72×10^{-15}	0.10×10^{-8}	0.36×10^{-10}	0.96×10^{-8}	0.20×10^{-30}

$$\frac{|\lambda'_{ex} - \lambda'_{ap}|}{\lambda'_{ex}} = \frac{|r'_{ex} - r'_{ap}|}{(\lambda - r) + r'_{ex}} < \frac{|r'_{ex} - r'_{ap}|}{r'_{ex}}. \tag{92}$$

It is typically even easier to estimate the throughput $\lambda - \lambda'$ accurately because typically $\lambda > 2\lambda'_{ex}$ so that

$$\frac{|(\lambda - \lambda'_{ex}) - (\lambda - \lambda'_{ap})|}{\lambda - \lambda'_{ex}} = \frac{|\lambda'_{ex} - \lambda'_{ap}|}{\lambda - \lambda'_{ex}} < \frac{|\lambda'_{ex} - \lambda'_{ap}|}{\lambda'_{ex}}. \tag{93}$$

Indeed, *for large capacities the throughput and the larger overflow rate are remarkably well approximated by the simple fluid approximation in (1).*

Table 5 shows that the direct Brownian approximations for the overflow rates are spectacularly good for the $E_2/E_2/1/11$ model, even better than for the M/M/1/11 model. Consequently, the M/M/1/C refinement, while not bad, is not an improvement. Neither is the GI/G/1/∞ refinement.

The two exact overflow rate values for ρ near 1 in the case $C = 100$ in Table 6 give strong evidence of the heavy-traffic limit. for the $0.8 < \rho < 1/0.8$, the exact overflow rates for the D/H$_2^b$/1/C and D/M^X/1/C models are quite close and the direct Brownian approximation is quite good. However, as ρ moves further away from 1 the difference between the D/M^X/1/C and D/H$_2^b$/1/C exact values grows. *In order to even roughly predict these small blocking probabilities when* ρ is not near 1 (e.g., when $\rho = 0.5$ or 2.0), *we evidently need more model detail than the arrival rate and the SCVs.* (However, the Brownian approximations perform significantly better than the M/M/1/C approximation, which ignores the SCVs.)

Table 7. A comparison of expressions for the overflow rates in the case of capacity $C = 10$ and $c_r^2 = 0$. Two job arrival SCVs are considered: $c_\lambda^2 = 4$ and 1. The job overflow rate λ' is displayed when $\lambda \leq r$, while the token overflow rate r' is dispayed when $r < \lambda$.

						Brownian approximations			
r	λ	c_λ^2	Ovfl. rate	D/H_2^b/1/C	D/M^X/1/C	Direct	M/M/1/C refinement	GI/G/1/∞ refinement	M/M/1/C
0.5	1.0	1		0.36×10^{-7}	0.36×10^{-7}	0.23×10^{-4}	0.86×10^{-5}	0.65×10^{-6}	0.00024
		4		0.0065	0.0019	0.045	0.0171	0.0032	0.00024
0.6	1.0	1		0.35×10^{-5}	0.35×10^{-5}	0.13×10^{-3}	0.72×10^{-4}	0.14×10^{-4}	0.00146
		4		0.0173	0.0093	0.063	0.0336	0.0125	0.00146
0.7	1.0	1	r'	0.000115	0.000115	0.00075	0.00050	0.000215	0.0061
		4		0.0358	0.0277	0.086	0.0575	0.0322	0.0061
0.8	1.0	1		0.00167	0.00167	0.0037	0.00289	0.00205	0.0188
		4		0.063	0.060	0.116	0.090	0.065	0.0188
0.9	1.0	1		0.0122	0.0122	0.0156	0.0134	0.0123	0.046
		4		0.100	0.108	0.154	0.132	0.115	0.046
1.0	1.0	1		0.0484	0.0484	0.050	0.0455	0.0455	0.091
		4		0.147	0.167	0.200	0.182	0.182	0.091
1.0	0.9	1		0.0133	0.0133	0.0122	0.0104	0.0143	0.046
		4		0.0893	0.1205	0.135	0.1148	0.107	0.046
1.0	0.8	1		0.00236	0.00236	0.00136	0.00105	0.00331	0.0188
		4		0.0731	0.0834	0.080	0.0621	0.0542	0.0188
1.0	0.7	1	λ'	0.00028	0.00028	0.000057	0.000038	0.00059	0.0061
		4		0.0168	0.0553	0.040	0.0266	0.0222	0.0061
1.0	0.6	1		0.23×10^{-4}	0.23×10^{-4}	0.65×10^{-6}	0.35×10^{-6}	0.84×10^{-4}	0.00146
		4		0.0038	0.035	0.0148	0.0079	0.0065	0.00146
1.0	0.5	1		0.115×10^{-5}	0.115×10^{-5}	0.10×10^{-8}	0.40×10^{-9}	0.87×10^{-5}	0.00024
		4		0.00042	0.021	0.0034	0.00128	0.00118	0.00024

For $\rho \leq 0.8$, the GI/G/1/∞ refinement performs remarkably poorly, which seems to be due to the very small numbers involved when $|(1 - \rho)C|$ is very large. On the other hand, the GI/G/1/∞ refinement seems to do quite well in Table 7 when $|(1 - \rho)C|$ is never large.

In summary, very small probabilities associated with large $|(1 - \rho)C|$ seem hard to estimate accurately. When $|(1 - \rho)C|$ is not large, the direct Brownian approximation provides a good rough approximation. The refinements sometimes help, but not consistently so. As a specific numerical procedure based on the Brownian approximation, we would suggest the GI/G/1/∞ refinement when $|(1 - \rho)C| \leq 10$. The Markov chain approximations in Section 5 seem better. However, note that the D/M^X/1/C and D/H_2^b/1/C values are not very good approximations for each other when $|(1 - \rho)C|$ is large. Thus, when $|(1 - \rho)C|$ is large, we would try to obtain and exploit additional model structure.

8.3 *Overflow SCVs*

Surprisingly, the most reliable Brownian approximations for the $D/H_2^b/1/C$ model are the approximations for the overflow SCVs. We begin to hope this will be so when we see how close the direct Brownian approximation in (29) is to the exact M/M/1/C value in (6) over the full range of ρ. Extensive numerical calculations indicate that the maximum percent error occurs at $C = 1$. At $C = 1$, the relative error is $(2C + 3)/(2C^2 + 4C + 3)$, which is approximately $1/(C + 1)$ when C is large. Further support is provided by the asymptotics in Section 6, which agree with the direct Brownian approximation when $|1 - \rho|$ gets large.

Tables 8 and 9 compare expressions for the SCVs partially characterizing the overflow processes in the cases of Tables 2 and 3. For the $D/M^X/1/C$ model we display the exact SCV c_T^2 for the renewal token overflow process, numerically computed as described in Section 5. For the $D/H_2^b/1/C$ model, we display simulation estimates. However, it is important to note that *the overflow processes are not renewal processes in this case.* We thus used simulation to estimate the *index of dispersion for intervals* (IDI) of the overflow process; see pp. 70–72 of Cox and Lewis [1966] and Section III of Fendick and Whitt [1989]. In particular, let S_n be the sum of n consecutive interoverflow intervals. Then the IDI is the function

$$I(n) = n \frac{\text{Var } (S_n)}{[E(S_n)]^2}, \qquad n \geq 1. \tag{94}$$

Each value of $I(n)$ is a candidate approximation for the SCV, because for a renewal process $I(n)$ is constant, equaling the SCV for all n.

For the $D/H_2^b/1/C$ model, we estimated the IDIs of the two overflow processes for $1 \leq n \leq 100$ using runs of 500,000 arrivals from a special purpose FORTRAN simulation of the token bank. In some cases, we display two values for each of c_T^2 and c_J^2. The first value is the estimate of $I(1)$, and the second value is the estimate of $I(100)$. In other cases (when ρ is not near 1), we display only the two values for the SCV of the dominant overflow process, because the other overflow process produced too few observations. Based on independent replications, we conclude that the halfwidths of the 90% confidence intervals for the stated values are less than 5% of the estimates.

Our first conclusion is that, even though the overflow processes are not exactly renewal processes, they are nearly so. The IDIs are nearly constant for the overflow processes, just as for the token arrival process in the $D/M^X/1/C$ model, which is actually a renewal process. This is in distinct contrast to the IDIs of the stream of admitted jobs; see Section 8.4.

Our second conclusion, in support of Theorems 3.1 and 4.1, is that the two overflow SCVs c_T^2 and c_J^2 tend to be nearly equal in the $D/H_2^b/1/C$ model. As ρ moves away from 1, the SCVs evidently become more different, but then the dominant stream is more important.

Table 8. A comparison of expressions for the SCVs of the overflow processes in the case of capacity $C = 100$ with $c_\lambda^2 = 10$ and $c_r^2 = 0$.

	D/H_2^b/1/C simulation			Brownian approximations			
$\rho = r/\lambda$	c_T^2	c_J^2	D/M^X/1/C exact c_T^2	Direct	M/M/1/C refinement	GI/G/1/∞ refinement	M/M/1/C
0.50		20.8–21.0	10.0	20.0	20.0	8.5	3.0
0.60		25.9–26.3	15.1	24.9	24.9	13.1	4.0
0.70	31.7–	34.2–34.9	23.3	32.5	32.5	21.0	5.7
0.80	54.7–32.5	47.2–46.8	38.0	44.3	44.3	36.3	9.0
0.90	66.0–66.9	62.4–62.6	57.6	58.9	58.9	58.9	19.0
0.95	66.7–69.7	70.8–63.8	64.0	64.5	64.7	65.6	37.1
1.00	71.6–69.7	65.7–60.1	65.0	66.7	67.3	67.3	67.3
1/0.95	66.4–64.3	62.3–60.2	60.9	64.3	64.4	65.4	37.1
1/0.90	61.9–60.0	62.9–43.4	53.2	57.4	57.4	57.3	19.0
1/0.80	40.3–39.3	43.5–	35.6	37.8	37.8	31.5	9.0
1/0.70	23.7–24.7		22.8	23.3	23.3	16.8	5.7
1/0.60	13.4–15.8		14.9	15.0	15.0	10.0	4.0
1/0.50	6.37–10.1		10.0	10.0	10.0	6.3	3.0

Table 9. A comparison of expressions for the SCVs of the overflow processes in the case of capacity $C = 10$ and $c_r^2 = 0$.

			D/H_2^b/1/C simulation			Brownian approximations				
r	λ	c_λ^2	c_T^2	c_J^2	D/M^X/1/C exact c_T^2	Direct	M/M/1/C refinement	GI/G/1/∞ refinement	§6	M/M/1/C
0.5	1.0	1		1.99–1.97	1.51	2.00	1.98	1.87	2.00	2.98
		4	5.4–6.1	7.3–7.5	3.95	5.53	5.58	3.99	8.00	2.98
0.6	1.0	1		2.48–2.46	1.96	2.49	2.55	2.32	2.50	3.87
		4	6.7–6.6	8.0–8.3	5.07	5.89	6.03	5.33	10.00	3.87
0.7	1.0	1	3.74–	3.31–3.26	2.73	3.25	3.40	3.08	3.33	4.99
		4	7.5–7.3	8.6–8.8	5.90	6.20	6.50	6.38	13.33	4.99
0.8	1.0	1		4.75–4.60	4.09	4.43	4.76	4.46	5.00	6.17
		4	8.1–8.2	8.9–9.0	6.29	6.45	6.94	6.95	20.00	6.17
0.9	1.0	1	5.6–5.8	6.4–6.4	5.84	5.89	6.46	6.38	10.00	7.05
		4	8.5–8.8	9.1–9.1	6.29	6.61	7.25	7.26	40.0	7.05
1.0	1.0	1	6.6–5.6	7.3–7.1	6.57	6.67	7.36	7.36	—	7.36
		4	8.7–9.1	9.0–8.9	6.04	6.67	7.36	7.36	—	7.36
1.0	0.9	1	5.5–5.0	6.6–5.8	5.46	5.74	6.29	6.47	9.00	7.05
		4	8.7–9.1	8.6–8.8	5.62	6.60	7.24	7.25	36.00	7.05
1.0	0.8	1	3.61–3.67	4.3–3.8	3.64	3.78	4.07	4.81	4.00	6.17
		4	8.0–8.2	7.8–7.4	5.04	6.33	6.82	6.86	16.00	6.17
1.0	0.7	1	2.28–2.27	2.13–	2.30	2.33	2.44	3.52	2.33	4.99
		4	6.8–7.7	6.8–6.1	4.34	5.79	6.07	6.07	9.33	4.99
1.0	0.6	1	1.49–1.47		1.50	1.50	1.53	2.74	1.50	3.87
		4	4.8–6.0	4.9–4.2	3.59	4.91	5.02	4.82	6.00	3.87
1.0	0.5	1	1.00–0.96		1.00	1.00	1.00	2.24	1.00	2.98
		4	2.9–4.1	3.0–	2.84	3.78	3.81	3.53	4.00	2.98

Tables 8 and 9 also contain the direct Brownian approximation, the two refinements, the M/M/1/C approximation and the approximation in Section 6 (the last only in Table 9). Our third conclusion is that the direct Brownian approximation does remarkably well, especially when we apply it to the dominant stream. (However, the case $c_\lambda^2 = 4$ in Table 9 is perhaps an exception. Note that the direct Brownian approximation is better than the $D/M^X/1/C$ exact value for c_T^2 as an approximation for c_J^2 and c_T^2 in the $D/H_2^b/1/C$ model though.) The M/M/1/C refinement differs very little from the direct approximation, while the GI/G/1/∞ refinement seems to produce a worse approximation. The approximation in Section 6 is also very accurate for ρ not near 1, but as $\rho \to 1$ the approximation in Section 6 deteriorates dramatically.

We conclude that the direct Brownian approximations of the overflow SCVs are sufficiently accurate for parametric-decomposition approximations as required for the multiclass throttle in Berger and Whitt [1992]. For the D/M/1/C dedicated banks in the multiclass throttle in Berger and Whitt [1992], we use the exact value of the token overflow SCV c_T^2 and the direct Brownian approximation for the job overflow SCV c_J^2. However, the Brownian approximation for c_T^2 seems to be as good as the exact value for engineering purposes. The D/M/1/C exact analysis seems most important for determining the overflow rates.

8.4 *The Accepted Job Stream*

In Section 6 we indicated that the accepted job stream should be nearly a renewal process with $c_A^2 \approx c_\lambda^2$ when $\rho > 1$, but the accepted job stream should not be nearly a renewal process when $\rho < 1$. When $\rho < 1$, we should have $c_A^2 \approx c_r^2$ using the asymptotic behavior of the arrival process, but the SCV of a stationary interval should reflect c_λ^2 as well as c_r^2. This analysis is substantiated by Table 10, which plots the IDI values $I(1)$ and $I(100)$ for the accepted job streams in the cases of Tables 3 and 4. For $\rho > 1$, we have $I(1) \approx I(100) \approx c_\lambda^2$. For $\rho < 1$, $I(100)$ is substantially less than $I(1)$.

Table 10. Values of the index of dispersion for intervals (IDI) for the accepted steam of jobs in the $D/H_2^b/1/C$ model estimated from simulation.

		$C = 10, c_\lambda^2 = 1$		$C = 10, c_\lambda^2 = 4$		$C = 100, c_\lambda^2 = 10$	
r	λ	$I(1)$	$I(100)$	$I(1)$	$I(100)$	$I(1)$	$I(100)$
0.5	1.0	0.37	0.0077	1.84	0.17	4.9	1.04
0.6	1.0	0.49	0.015	2.21	0.33	6.0	1.95
0.7	1.0	0.61	0.033	2.55	0.57	7.0	3.3
0.8	1.0	0.73	0.14	2.85	0.89	8.0	5.1
0.9	1.0	0.85	0.20	3.12	1.27	8.9	7.0
1.0	1.0	0.93	0.44	3.35	1.71	9.6	8.8
1.0	0.9	0.98	0.74	3.56	2.23	9.9	9.9
1.0	0.8	1.00	0.92	3.75	2.84	10.0	10.3
1.0	0.7	1.00	0.97	3.89	3.53	10.1	10.4
1.0	0.6	1.00	0.98	3.97	3.96	10.0	10.1
1.0	0.5	1.00	0.96	4.01	4.12	10.1	10.3

8.5. A Refinement to the Brownian Approximation for Discreteness

In (88) we noted that the direct Brownian approximation for the mean in the M/M/1/∞ model is off by exactly 1/2. Looking at the M/M/1/C formulas in (3)–(6) and the corresponding RBM formulas in (23)–(29), we see that C tends to appear in the RBM formula where $C + 1$ appears in the corresponding M/M/1/C formula. It turns out that both these discrepancies can be removed by introducing a refinement to account for the discreteness. In particular, motivated by continuous-distribution approximations for discrete distributions, we suggest approximating the M/M/1/C model, by RBM with barriers at $-1/2$ and $C + 1/2$ instead of at 0 and C. We then approximate the steady-state pmf value $p(k)$ by the integral $\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} p(x)\,dx$.

Of course, this discreteness refinement could be applied more generally, but it obviously could make the approximate mean negative in some cases. However, at least for the M/M/1/C model this refinement is effective. With this discreteness refinement, the approximate mean for the M/M/1/∞ model is shifted by 1/2, so that the error in (88) is eliminated completely. Moreover, the distance between the two barriers for RBM is now $C + 1$ instead of C, so that $C + 1$ tends to appear in the new formulas instead of C.

The improvement provided by this discreteness refinement is dramatically demonstrated in the case $\rho = 1$. Since

$$\frac{1}{C} - \frac{1}{C+1} = \frac{1}{C(C+1)},$$

there are errors of $1/C(C + 1)$ in the RBM approximations for the M/M/1/C pmf in (3) and the overflow probabilities in (5) which are eliminated by the refinement, while the RBM approximation for steady-state mean remains exact. The error for the overflow SCV in (6) when $\rho = 1$ is reduced from $2/3 + 1/3(C + 1)$ to $1/3(C + 1)$.

We can also show that the refinement leads to improved approximations when $\rho \neq 1$. Let λ'_{RBM} and r'_{RBM} be the direct RBM approximations in (25) and (26); and let λ'_{RBM^*} and r'_{RBM^*} be the refinement for discreteness with $(C + 1)$ substituted for C.

THEOREM 8.1. Consider the M/M/1/C model for which $\theta = -2(1 - \rho)/(1 + \rho)$.

(a) If $\rho < 1$, then

$$r' \equiv \frac{\lambda - r}{\rho^{-(C+1)} - 1} < r'_{\text{RBM}^*} \equiv \frac{\lambda - r}{e^{-\theta(C+1)} - 1} < r'_{\text{RBM}} \equiv \frac{\lambda - r}{e^{-\theta C} - 1}.$$

(b) If $\rho > 1$, then

$$\lambda' \equiv \frac{r - \lambda}{\rho^{(C+1)} - 1} < \lambda'_{\text{RBM}^*} \equiv \frac{r - \lambda}{e^{\theta(C+1)} - 1} < \lambda'_{\text{RBM}} \equiv \frac{r - \lambda}{e^{\theta C} - 1}.$$

Proof. We only treat (a) since the argument for (b) is similar. It suffices to show that $-\log \rho > -\theta$, but

$$-\log \rho = -\log (1 - (1 - \rho)) = \sum_{k=1}^{\infty} \frac{(1 - \rho)^k}{k}, \tag{95}$$

while

$$-\theta = \frac{2(1 - \rho)}{1 + \rho} = \frac{2(1 - \rho)}{2 - (1 - \rho)} = \sum_{k=1}^{\infty} \frac{(1 - \rho)^k}{2^{k-1}}. \quad \blacksquare \tag{96}$$

From (95) and (96), we can also derive expressions for the asymptotic errors as $\rho \to 1$, e.g., we obtain

$$-\log \rho - (-\theta) = \frac{(1 - \rho)^3}{12} + \frac{(1 - \rho)^4}{8} + \frac{11(1 - \rho)^5}{80} + O(1 - \rho)^6. \tag{97}$$

We apply (95)–(97) to obtain the following result.

THEOREM 8.2. When $\rho < 1$, the relative error is

$$\frac{|r' - r'_{\text{RBM}^*}|}{r'} = \frac{e^{-\theta(C+1)} - e^{-(C+1)\log \rho}}{e^{-\theta(C+1)} - 1} = \frac{(1 - \rho)^2}{12} + O(1 - \rho)^3.$$

The asymptotic behavior in Theorem 8.2 also applies to the difference between the means in the M/M/1/∞ and M/M/1/C models (i.e., if we omit the $\rho/(1 - \rho)$ term in (4) when $\rho \neq 1$.

9. Insights from the Brownian Overflow Rate Formula

Even though the direct Brownian approximation for the job overflow rate in (25) is not especially accurate when $|(1 - \rho)C|$ is large, we contend that this simple Brownian formula can provide important insights for system design (especially when $|(1 - \rho)C|$ is indeed not large). We consider the case $c_r^2 = 0$, which occurs in standard token banks. (A similar analysis can be done for $c_r^2 > 0$.) An application of particular interest is the policing function in ATM networks, where (25) could be used to provide guidance in setting parameters of the token bank to be appropriate for the service contract with the end user and to avoid falsely marking or dropping cells. In this application, relevant values of ρ are greater than 1. Equation (25) can also be used to provide insight into the marking or dropping rates when the user violates the service contract, in which case ρ would be less than 1.

Also the Brownian approximation (25) with $c_r^2 = 0$ can be applied to finite waiting room queues with constant service times as in, for example, output buffers for fixed-length ATM cells. To represent ATM cell buffers in our token bank framework we are interested

in the dual queue in which r plays the role of service rate and λ plays the role of the arrival rate. For the ATM cell buffer, we are primarily interested in the overflow probability λ'/λ when $\rho > 1$.

Suppose that it has been stipulated that the long-run job overflow probability λ'/λ should be $e^{-\gamma}$. We want to know how the capacity C and the ratio $\rho = r/\lambda$ should depend on the overflow rate exponent γ and the arrival process characteristics λ and c_λ^2. Note that for $\rho < 1$, λ'/λ cannot be made arbitrarily small and thus γ cannot be chosen at will. The smallest blocking of jobs occurs when $C = \infty$, in which case $\lambda' = \lambda - r$ and $\lambda'/\lambda = 1 - \rho$. Thus, for $\rho < 1$ and finite C, $\lambda'/\lambda > 1 - \rho$, and the choice for γ is restricted by $(1 - \rho)e^{\gamma} < 1$. From (25), noting that $\theta = 2(\rho - 1)/c_\lambda^2$ since $c_r^2 = 0$, we see that

$$C = \begin{cases} \theta^{-1} \log\,(1 + e^{\gamma}(\rho - 1)) = c_\lambda^2 \dfrac{\log(1 + e^{\gamma}(\rho - 1))}{2(\rho - 1)}, & \rho \neq 1, \\ c_\lambda^2\, e^{\gamma}/2, & \rho = 1. \end{cases} \tag{98}$$

From (98), we see that *the required capacity C is directly proportional to the arrival SCV c_λ^2*. This seems to be a valuable first-order approximation, but unfortunately for complex job arrival processes as are anticipated in the ATM environment, this leaves open the very difficult problem of finding an appropriate variability parameter c_λ^2. The utility of (98) is not great when the appropriate variability parameter c_λ^2 depends critically on the other parameters. Nevertheless, it may be possible to estimate an appropriate SCV c_λ^2 by measurement, as discussed in Fendick and Whitt [1989]. At least, (98) should provide a useful rough check with more sophisticated procedures.

To see how (98) works when c_λ^2 is unambiguously well defined, consider the D/M/1/C model in which $c_\lambda^2 = 1$. Suppose that $e^{-\gamma} = 10^{-6}$ and $\rho = 1.1$. Then (98) predicts that the required capacity is 58. At this capacity, the actual blocking probability is 1.76×10^{-6}. The actual required capacity to have $\lambda'/\lambda \leq 10^{-6}$ is $C = 62$.

If in the last example we increase c_λ^2 from 1 to 2, then (98) predicts that the required capacity should double; i.e., we need $C = 116$. For the $D/M^X/1/C$ model with $c_\lambda^2 = 2$, $\lambda'/\lambda = 2.2 \times 10^{-6}$ at $C = 116$, and λ'/λ dips below 10^{-6} at $C = 125$. Although these approximations are not remarkably accurate (and they get worse as ρ increases), we think they have potential for providing practical engineering guidelines.

For the ATM environment, we anticipate that the stipulated overflow rate $e^{-\gamma}$ will be small and ρ will not be too close to 1, so that $e^{-\gamma}(\rho - 1) \gg 1$, where $\gg$ denotes "much greater than." When $e^{\gamma}(\rho - 1) \gg 1$, we can ignore the 1 in the logarithm, and obtain

$$C \approx c_\lambda^2 \left\lceil \frac{\log(\rho - 1) + \gamma}{2(\rho - 1)} \right\rceil \quad \text{for } \rho > 1. \tag{99}$$

which implies that *C is approximately a positive linear function of the overflow rate exponent γ*. For example, if $\rho = 2$, then $C = c_\lambda^2\, \gamma/2$.

From (99) we obtain a convenient rougher approximation if we assume that $\gamma + \log(\rho - 1) \approx \gamma$. Then

$$C \approx \frac{c_\lambda^2 \gamma}{2(\rho - 1)} \quad \text{or} \quad \theta C \approx \gamma \qquad \text{for } \rho > 1. \tag{100}$$

For example, if $e^{-\gamma} = 10^{-6}$ and $1.5 \leq \rho \leq 3$, then $\gamma = 13.8$ and $|\log(\rho - 1)| \leq 0.69$, so that the simplification is roughly valid.

We can also view (98) as an expression for the required capacity C as a function of ρ. With γ and c_λ^2 held fixed, the required capacity C typically (but not always) decreases as ρ increases. In particular, regarding C as a function of ρ, we can apply (98) to quantify this relation, obtaining the following result. For $\rho \neq 1$,

$$\begin{aligned} \frac{\partial C}{\partial \rho} &= \frac{1}{2(\rho - 1)} \left[\frac{c_\lambda^2 e^\gamma}{1 + e^\gamma(\rho - 1)} - 2C \right] \\ &= \frac{1}{\theta(\rho - 1)} [1 - e^{-\theta C} - \theta C]. \end{aligned} \tag{101}$$

To justify (101), take the partial derivatives with respect to ρ in (98) to obtain

$$\frac{\partial C}{\partial \rho} = \frac{c_\lambda^2 e^\gamma}{2(\rho - 1)(1 + e^\gamma(\rho - 1))} - \frac{c_\lambda^2 \log(1 + e^\gamma(\rho - 1))}{2(\rho - 1)^2}. \tag{102}$$

From (98), the second term in (102) is $C/(\rho - 1)$, and the first expression in (101) follows:

$$\begin{aligned} \frac{\partial C}{\partial \rho} &= \frac{c_\lambda^2 e^\gamma}{2(\rho - 1)(1 + e^\gamma(\rho - 1))} - \frac{C}{\rho - 1} \\ &= \frac{1}{\rho - 1} \left[\frac{c_\lambda^2 e^\gamma}{2(1 + e^\gamma(\rho - 1))} - C \right]. \end{aligned}$$

From (25), we see that

$$e^{-\gamma} \equiv \frac{\lambda'}{\lambda} = \frac{\rho - 1}{e^{\theta C} - 1},$$

so that

$$\begin{aligned} \frac{\partial C}{\partial \rho} &= \frac{1}{\rho - 1} \left[\frac{c_\lambda^2 (e^{\theta C} - 1)}{2(\rho - 1) e^{\theta C}} - C \right] \\ &= \frac{1}{\rho - 1} \left[\frac{e^{\theta C} - 1}{\theta e^{\theta C}} - C \right] \end{aligned}$$

$$= \frac{1}{\theta(\rho - 1)} [1 - e^{-\theta C} - \theta C].$$

If $\theta C \gg 1$ (which implies $\rho > 1$) then (101) leads to the simple approximation

$$\frac{\partial C}{\partial \rho} \approx \frac{1}{\theta(\rho - 1)} (-\theta C) = -\frac{C}{\rho - 1}.$$

To see how ρ should depend on c_λ^2 for fixed λ and C, regard ρ as a function of c_λ^2 and differentiate with respect to c_λ^2 in (98) to obtain

$$\frac{\partial \rho}{\partial c_\lambda^2} = \frac{\rho - 1}{c_\lambda^2 [1 - c_\lambda^2 e^{(\gamma - \theta C)}/2C]}. \tag{103}$$

The simple approximation

$$\frac{\partial \rho}{\partial c_\lambda^2} \approx \frac{\gamma}{2C} \tag{104}$$

obtained from (100) also comes from (103) upon making further simplifying assumptions.

We can also use (25) to obtain insight into the *sensitivity* of λ' and r' to small changes in the parameter σ^2, C, and μ. For the case of $\rho \neq 1$, by differentiating, we obtain

$$\frac{\partial \lambda'}{\partial \sigma^2} = \frac{\partial r'}{\partial \sigma^2} = \frac{\mu \theta C}{\sigma^2} \frac{e^{\theta C}}{(e^{\theta C} - 1)^2} = -\frac{C}{\sigma^2} \left(\frac{\partial \lambda'}{\partial C} \right) = -\frac{C}{\sigma^2} \left(\frac{\partial r'}{\partial C} \right), \tag{105}$$

which implies a simple relation between the *elasticities* of λ' with respect to σ^2 and C, namely,

$$e(\lambda', \sigma^2) \equiv \left(\frac{\partial \lambda'}{\partial \sigma^2} \right) \left(\frac{\sigma^2}{\lambda'} \right) = \frac{\theta C e^{\theta C}}{e^{\theta C} - 1} = -\left(\frac{\partial \lambda'}{\partial C} \right) \left(\frac{C}{\lambda'} \right) \equiv -e(\lambda', C). \tag{106}$$

Similarly,

$$e(\lambda', \mu) \equiv \left(\frac{\partial \lambda'}{\partial \mu} \right) \left(\frac{\mu}{\lambda'} \right) = 1 - \frac{\mu \theta C e^{\theta C}}{\lambda'(e^{\theta C} - 1)^2} = 1 - e(\lambda', \sigma^2). \tag{107}$$

Roughly speaking, a small percentage increase in σ^2 leads to $e(\lambda', \sigma^2)$ times that percentage increase in λ'. For the case $\rho = 1$, $\lambda' = \sigma^2/2C$ in (25) and

$$e(\lambda', \sigma^2) = -e(\lambda', C) = 1 \quad \text{and} \quad e(\lambda', \mu) = 0. \tag{108}$$

For example, suppose that $\lambda = 1.0$, $r = 1.1$, $c_r^2 = 0$, $c_\lambda^2 = 10$, and $C = 100$. Then $\rho = 1.1$, $\mu = 0.1$, $\sigma^2 = 10$, $\theta = 0.02$, $\theta C = 2.0$, $\lambda' = 0.0157$, $\partial\lambda'/\partial\sigma^2 = 0.00362$, $e(\lambda',\sigma^2) = 2.313$, and $e(\lambda', \mu) = -1.313$. In this case we see that a 1% increase in μ (or an equivalent increase in r) leads to approximately a 1.3% decrease in λ', while a 1% increase in σ^2 (or, equivalently, c_λ^2) leads to approximately a 2.3% increase in λ'.

10. Conclusions

In this paper we have obtained some useful insights into both rate-control throttles and Brownian approximations. We compared the token-bank throttle to the leaky-bucket throttle in Section 7, and in the rest of the paper developed approximations for both, giving special attention to the token-bank throttle. When the capacity of the token bank is not too small and the job arrival rate differs from the token arrival rate, the throughout and overflow rates are surprisingly well described by the simple fluid model in Section 2. The Poisson approximation in Section 3 gives some idea of the stochastic effect, but the other stochastic approximations seem far better.

To describe the steady-state distribution of the number of tokens in the token bank and to describe the overflow processes, the Brownian approximation is quite effective. The Brownian approximation is about as accurate an approximation for the $D/H_2^b/1/C$ and $D/M^X/1/C$ models as each is as for the other, but of course the Brownian approximation involves relatively simple closed-form formulas instead of rather involved Markov chain computations. The simple direct Brownian formulas in Section 4.2 are appealing for their simplicity, and should be useful for many engineering purposes, as illustrated by the sensitivity analysis associated with blocking-probability formula (25) in Section 9. The relatively simple refinements to the Brownian approximation in Sections 4.3 and 4.4 are effective for obtaining improved numerical accuracy in the case of the mean number in system and possibly for the overflow rates. The discreteness refinement in Section 8.5 also seems promising.

Our numerical comparisons have provided a useful perspective on the accuracy of the Brownian approximations. Consistent with previous experience, we find that the accuracy is often good, but sometimes it is not. For example, the quality of the direct Brownian approximations for the steady-state mean and the overflow rates for the $E_2/E_2/1/11$ queue in Section 8 are consistent with the 5% relative error reported by Dai and Harrison [1991], but the quality of the direct Brownian approximation for the $D/M/1/\infty$ queue with small ρ are not. We have suggested that the accuracy of the RBM approximation for the logarithm of the blocking probability might be comparable to the accuracy for the mean. We have also suggested that $|(1 - \rho)C|$ might be used to estimate the accuracy of the Brownian approximation for the blocking probabilities.

The Brownian approximation is also effective for determining the SCVs partially characterizing two overflow streams. Unlike for the overflow rates, for the overflow SCVs the direct Brownian approximation is quite accurate even when ρ is not near 1 (e.g., $\rho = 0.5$). An alternative approach for the overflow SCVs is to use the exact SCV for the token arrival process in the $D/M^X/1/C$ model for both jobs and tokens, which is supported by the Brownian analysis. However, for the $D/H_2^b/1/C$ model, the Brownian approximation for the overflow SCVs seems to be just as good. The direct Brownian approximations

for the SCVs of the overflow processes seem suitable for engineering applications, and are being applied to analyze the multiclass throttle in Berger and Whitt [1992]. For the overflow processes in the $D/H_2/1/C$ and $D/M^X/1/C$ token banks, we suggest using renewal-process approximations based on the exact overflow rates computed as described in Section 5 and the Brownian SCVs in (29).

Acknowledgment.

We are grateful to Kerry Fendick and Marty Reiman for helpful discussions about the Brownian approximation. We are grateful to Ruth Williams for providing the direct proof of Theorem 4.1 (appearing in her paper).

References

Abate, J., and Whitt, W., 1988. Transient behavior of the M/M/1 queue via Laplace transforms. *Adv. Appl. Prob.*, 20:145–178.

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.

Berger, A.W., 1991a. Overload control using a rate control throttle: Selecting token bank capacity for robustness to arrival rates. *IEEE Trans. Aut. Cont.*, 36: 216–219. (Also *Proc. 28th IEEE Conf. on Decision and Control*, 1989, 2527–2529).

Berger, A.W., 1991b. Performance analysis of a rate-control throttle where tokens and jobs queue. *IEEE J. Sel. Areas Commun.*, 9:165–170. (also *IEEE INFOCOMM*, June 1990, 30–38).

Berger, A.W., and Whitt, W., 1990. A multi-class input-regulation throttle. *Proc. 29th IEEE Conf. on Decision and Control*, 2106–2111.

Berger, A.W., and Whitt, W., 1992. A multi-class rate-control throttle. In preparation.

Billingsley, P., 1968. *Convergence of Probability Measures.* New York: Wiley.

Borovkov, A.A., 1965. Some limit theorems in the theory of mass service. II. *Theor. Prob. Appl.*, 10:375–400.

Borovkov, A.A., 1976. *Stochastic Processes in Queueing Theory.* New York: Springer-Verlag.

Budka, K.C., 1990. First- and Second-Order Stochastic Properties of Rate-Based Flow Control Mechanisms. Department of Industrial Engineering and Operations Research, Columbia University.

Budka, K.C., and Yao, D.D., 1990. Monotonicity and Convexity Properties of Rate Control Throttles. Department of Industrial Engineering and Operations Research, Columbia University.

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

Chen, H., and Mandelbaum, A., 1991b. Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *Ann Probab.*, 19: 1463–1519.

Chen, H., and Whitt, W., 1991. Diffusion approximations for open queueing networks with service interruptions. AT&T Bell Laboratories, Murray Hill, NJ.

Coffman, E.G., Jr., Puhalsky, A.A., and Reiman, M.I., 1991. Storage-limited queues in heavy traffic. *Prob. Engr. Inf. Sci.*, 5: 499–522.

Coffman, E.G., Jr., and Reiman, M.I., 1984. Diffusion approximation for computer communications systems. pp. 33–53 in *Mathematical Computer Performance and Reliability*; Iazcolla, G., Courtois, P.J. and Hordijk, A. (eds.) Amsterdam: The North-Holland.

Cox, D.R., and Lewis, P.A.W., 1966. *The Statistical Analysis of Series of Events.* London: Metheun.

Csörgo', M., Horváth, L., and Steinebach, J., 1987. Invariance principles for renewal processes. *Ann. Probab*, 15:1441–1460.

Dai, J.G., and Harrison, J.M., 1991. Steady-state analysis of RBM in a rectangle: Numerical methods and a queueing application. *Ann. Appl. Prob.*, 1:16–35.

Doshi, B.T., and Heffes, H., 1983. Analysis of overload control schemes for a class of distributed switching machines. *10th Int. Teletraffic Congress*, Montreal, Canada, Paper No. 5.2.2.

Dupuis, P., and Ishii, H., 1991. On when the solution to the Skorokhod problem is Lipschitz continuous with applications. *Stochastics*, 35:31–62.

Eckberg, A.E., Luan, D.T., and Lucantoni, D.M., 1989. Bandwidth management: A congestion control strategy for broadband packet networks characterizing the throughput-burstiness filter. *Int., Teletraffic Congress Specialist Seminar*, Adelaide, Australia, (Sept.), Paper No. 4.4.

Eisenberg, M., 1983. A strict priority queueing system with overload control. *10th Int. Teletraffic Congress*, Montreal, Paper No. 1.3.2.

Elwalid, A., 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.

Fendick, K.W., and Rodrigues, M., 1991. A heavy-traffic comparison of shared and segregated buffer schemes for queues with the head-of-line processor-sharing discipline. *Queueing Systems*, 9: 163–190.

Fendick, K.W., and Whitt, W., 1989. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proc. IEEE*, 77:171–194.

Gaver, D.P., and Shedler, G.S.,1973a. Processor utilization in multiprogramming systems via diffusion approximations. *Oper. Res.*, 21:569–576.

Gaver, D.P., and Shedler, G.S., 1973b. Approximate models for processor utilization in multiprogrammed computer systems. *SIAM J. Comput.*, 2:183–192.

Gelenbe, E., 1975. On approximate computer system models. *J. ACM.*, 22: 261–269.

Gelenbe, E., and Mitrani, I., 1980. *Analysis and Synthesis of Computer Systems*. New York: Academic Press.

Glynn, P.W., and Whitt, W., 1987. Sufficient conditions for functional-limit-theorem versions of $L=\lambda W$. *Queueing Systems*, 1:279–287.

Harrison, J.M., 1985. *Brownian Motion and Stochastic Flow Systems*. New York: Wiley

Heyman, D.P., and Stidham, S., Jr., 1980. The relation between customer and time averages in queues. *Oper. Res.*, 28:983–994.

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

Iglehart, D.L., and Whitt, W., 1970b. Multiple channel queues in heavy traffic. II: Sequences, networks and batches. *Adv. Appl. Prob.*, 2:355–369.

Kalashnikov, V.V., and Rachev, S.T., 1990. *Mathematical Methods for Construction of Queueing Models*. Belmont, CA.: Wadsworth and Brooks/Cole.

Keilson, J., 1979. *Markov Chain Models—Rarity and Exponentiality*. New York: Springer-Verlag.

Kemeny, J.G., and Snell, J.L., 1959. Finite Markov Chains. Princeton: Van Nostrand.

Kennedy, D.P., 1973. Limit theorems for finite dams. *Stoch. Proc. Appl.* 1:269–278.

Kimura, T., 1985. Refining diffusion approximations for GI/G/1 queues: A tight discretization method. *Teletraffic Issues in an Advanced Information Society, ITC*-11. M. Akiyama (ed.), Amsterdam: Elsevier, 317–323.

Kimura, T., Ohno, K., and Mine, H., 1979. Diffusion approximation for the GI/G/1 queueing systems with finite capacity. II: The stationary behavior. *J. Oper. Res. Soc. Japan*, 22:301–319.

Klincewicz, J.G., and Whitt, W., 1984. On approximations for queues. II: Shape constraints. *AT&T Bell Lab. Tech. J.*, 63:139–161.

Kraemer, W., and Langenbach-Belz, M., 1976. Approximate formulae for the delay in the queueing system GI/G/1. *Proc. Eighth Int. Teletraffic Cong.*, Melbourne, 235-1/8.

Kroner, H., Theimer, T.H., and Briem, U., 1990. Queueing models for ATM systems—A comparison. *7th Teletraffic Congress Seminar Broadband Technologies*, (Oct.), Morristown, NJ.

Kühn, P., 1976. *Tables on Delay Systems*. Institute of Switching and Data Technics, University of Stuttgart.

Neuts, M.F., 1986. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore: Johns Hopkins University Press.

Newell, G.F., 1982. *Applications of Queueing Theory*, 2nd ed., London: Chapman and Hall.

Rathgeb, E.P., 1990. Policing mechanisms for ATM networks—Modelling Performance comparison. *7th Int. Teletraffic Congress Seminar, Broadband Technologies*. (Oct.), Morristown, NJ.

Reiman, M.I., 1984. Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9:441–458.

Ross, S., 1982. *Stochastic Processes*. New York: Wiley.

Seelen, L.P., Tijms, H.C., and van Hoorn, M.H., 1985. *Tables for Multi-Server Queues*. Amsterdam: North-Holland.

Sidi, M., Liu, W.Z., Cidon, I., and Gopal, I., 1989. Congestion control through input rate regulation. *GLOBECOM '89*, (Nov.) 1764–1768, Dallas, TX.

Sohraby, K., and Sidi, M., 1990. On the performance of bursty and corrleated sources subject to leaky bucket rate-based access control schemes. *IEEE INFOCOM '91*, (April), Bal Harbour, Florida, 426–434.

Sriram, K., and Whitt, W. 1986. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. Sel. Areas Commun.*, SAC4:833–846.

Sweet, A.L., and Hardin, J.C., 1970. Solutions for some diffusion processes with two barriers. *J. Appl. Prob.*, 7:423–431.

Turner, J. 1986. New directions in communications (or which way to the information age?). *IEEE Comm. Mag.*, 24:8–15.

Whitt, W., 1969. *Weak Convergence Theorems for Queues in Heavy Traffic.* Ph.D. thesis, Cornell University.

Whitt, W., 1974. Preservation of rates of convergence under mapping. *Z. Wahrsch. Gebiete*, 29:39–44.

Whitt, W., 1980. Some useful functions for functional limit theorems. *Math. Oper. Res.*, 5:67–85.

Whitt, W., 1982a. Approximating a point process by a renewal process. I: Two basic methods. *Oper. Res.*, 30:125–147.

Whitt, W., 1982b. Refining diffusion approximations for queues. *Oper. Res. Lett.*, 1:165–169.

Whitt, W., 1983. The queueing network analyzer. *Bell System Tech. J.*, 62:2779–2815.

Whitt, W., 1984a. On approximations for queues. I: Extremal distributions. *AT&T Bell Lab. Tech. J.*, 63:115–138.

Whitt, W., 1984b. On approximations for queues. III: Mixtures of exponential distributions. *AT&T Bell Lab. Tech. J.*, 63:163–175.

Whitt, W., 1985. Approximation for the GI/G/m queues. AT&T Bell Laboratories.

Whitt, W., 1989. An interpolation approximation for the mean workload in a GI/G/1 queue. *Oper. Res.*, 37:936–952.

Williams, R.J., 1991. Asymptotic variance parameters for the boundary local times of reflected brownian motion on a compact interval. Department of Mathematics, University of California at San Diego. To appear in *J. Appl. Prob.*

Wolff, R.W., 1982. Poisson arrival see time averages. *Oper. Res.*, 30:223–231.

Yao, D.D., and Buzacott, J.A., 1985a. Queueing models for a flexible machining station. Part I. Diffusion approximations. *Eur. J. Oper. Res.*, 19:233–241.

Yao, D.D., and Buzacott, J.A., 1985b. Queueing models for a flexible machining station. Part II: The method of Coxian phases. *Eur. J. Oper. Res.*, 19:242–252.