Chapter 5

Heavy-Traffic Limits for Fluid Queues

5.1. Introduction

In this chapter we see how the continuous-mapping approach can be applied to establish heavy-traffic stochastic-process limits for queueing models, and how those heavy-traffic stochastic-process limits, in turn, can be applied to obtain approximations for queueing processes and gain insight into queueing performance.

To establish the heavy-traffic stochastic-process limits, the general idea is to represent the queueing "content" process of interest as a reflection of a corresponding net-input process. For single queues with unlimited storage capacity, a one-sided one-dimensional reflection map is used; for single queues with finite storage capacity, a two-sided one-dimensional reflection map is used. These one-dimensional reflection maps are continuous as maps from D to D with all the principal topologies considered by virtue of results in Sections 13.5 and 14.8. Hence, FCLT's for scaled net-input processes translate into corresponding FCLT's for scaled queueing processes.

Thus we see that the relatively tractable heavy-traffic approximations can be regarded as further instances of the statistical regularity stemming from the FCLT's in Chapter 4. The FCLT for the scaled net-input processes may be based on Donsker's theorem in Section 4.3 and involve convergence to Brownian motion; then the limit process for the scaled queueing processes is reflected Brownian motion (RBM). Alternatively, the FCLT for the scaled net-input processes may be based on one of the other FCLT's in Sections

4.5 – 4.7 and involve convergence to a different limit process; then the limit process for the scaled queueing processes is the reflected version of that other limit process.

For example, when the net-input process can be constructed from partial sums of IID random variables with heavy-tailed distributions, Section 4.5 implies that the scaled net-input processes converge to a stable Lévy motion; then the limit process for the queueing processes is a reflected stable Lévy motion. The reflected stable Lévy motion heavy-traffic limit describes the effect of the extra burstiness due to the heavy-tailed distributions.

As indicated in Section 4.6, it is also possible to have more burstiness due to strong positive dependence or less burstiness due to strong negative dependence. When the net-input process has such strong dependence with light-tailed distributions, the scaled net-input processes may converge to fractional Brownian motion; then the limit process for the scaled queueing processes is reflected fractional Brownian motion.

In this chapter, attention will be focused on the "classical" Brownian approximation involving RBM and its application. For example, in Section 5.8 we show how the heavy-traffic stochastic-process limit with convergence to RBM can be used to help plan queueing simulations, i.e., to estimate the required run length to achieve desired statistical precision, as a function of model parameters. Reflected stable Lévy motion will be discussed in Sections 8.5 and 9.7, while reflected fractional Brownian motion will be discussed in Sections 8.7 and 8.8.

In simple cases, the continuous-mapping approach applies directly. In other cases, the required argument is somewhat more complicated. A specific simple case is the discrete-time queueing model in Section 2.3. In that case, the continuous-mapping argument applies directly: FCLT's for the partial sums of inputs V_k translate immediately into associated FCLT's for the workload (or buffer-content) process $\{W_k\}$, exploiting the continuity of the two-sided reflection map. The continuous-mapping approach applies directly because, as indicated in (3.5) in Chapter 1, the scaled workload process is exactly the reflection of the scaled net-input process, which itself is a scaled partial-sum process. Thus all the stochastic-process limits in Chapter 4 translate into corresponding heavy-traffic stochastic-process limits for the workload process in Section 2.3.

In this chapter we see how the continuous-mapping approach works with related continuous-time fluid-queue models. We start considering fluid queues, instead of standard queues (which we consider in Chapter 9), because fluid queues are easier to analyze and because fluid queues tend to serve as initial "rough-cut" models for a large class of queueing systems.

The fluid-queue models have recently become popular because of applications to communication networks, but they have a long history. In the earlier literature they are usually called dams or stochastic storage models; see Moran (1959) and Prabhu (1998). In addition to queues, they have application to inventory and risk phenomena.

In this chapter we give proofs for the theorems, but the emphasis is on the statement and applied significance of the theorems. The proofs illustrate the continuous-mapping approach for establishing stochastic-process limits, exploiting the useful functions introduced in Section 3.5. Since the proofs draw on material from later chapters, upon first reading it should suffice to focus, first, on the theorem statements and their applied significance and, second, on the general flow of the argument in the proofs.

5.2. A General Fluid-Queue Model

In a fluid-queue model, a divisible commodity (fluid) arrives at a storage facility where it is stored in a buffer and gradually released. We consider an *open model* in which fluid arrives *exogenously* (from outside). For such open fluid-queue models, we describe the buffer content over time. In contrast, in a standard queueing model, which we consider in Chapter 9, individual customers (or jobs) arrive at a service facility, possibly wait, then receive service and depart. For such models, we count the number of customers in the system and describe the experience of individual customers. The fluid queue model can be used to represent the unfinished work in a standard queueing model. Then the input consists of the customer service requirements at their arrival epochs. And the unfinished work declines at unit rate as service is provided.

In considering fluid-queue models, we are motivated to a large extent by the need to analyze the performance of evolving communication networks. Since data carried by these networks are packaged in many small packets, it is natural to model the flow as fluid, i.e., to think of the flow coming continuously over time at a random rate. A congestion point in the network such as a switch or router can be regarded as a queue (dam or stochastic storage model), where input is processed at constant or variable rate (the available bandwidth). Thus, we are motivated to consider fluid queues. However, we should point out that other approaches besides queueing analysis are often required to engineer communication networks; to gain perspective, see Feldmann et al. (2000, 2001) and Krishnamurthy and Rexford (2001).

5.2.1. Input and Available-Processing Processes

In this section we consider a very general model: We consider a single fluid queue with general input and available-processing (or service) processes. For any $t > 0$, let $C(t)$ be the cumulative input of fluid over the interval $[0, t]$ and let $S(t)$ be the cumulative available processing over the interval $[0, t]$. If there is always fluid to process during the interval $[0, t]$, then the quantity processed during $[0, t]$ is $S(t)$. We assume that $\{C(t) : t \geq 0\}$ and $\{S(t) : t \geq 0\}$ are real-valued stochastic processes with nondecreasing nonnegative right-continuous sample paths. But at this point we make no further structural or stochastic assumptions.

A common case is processing at a constant rate μ whenever there is fluid to process; then

$$S(t) = \mu t, \quad t \geq 0 \ . \tag{2.1}$$

More generally, we could have input and output at random rates. Then

$$C(t) = \int_0^t R_i(s)ds \quad \text{and} \quad S(t) = \int_0^t R_o(s)ds, \quad t \geq 0 \ , \tag{2.2}$$

where $\{R_i(t) : t \geq 0\}$ and $\{R_o(t) : t \geq 0\}$ are nonnegative real-valued stochastic processes with sample paths in D. For example, it is natural to have maximum possible input and processing rates ν_i and ν_o. Then, in addition to (2.2), we would assume that

$$0 \leq R_i(t) \leq \nu_i \quad \text{and} \quad 0 \leq R_o(t) \leq \nu_o \quad \text{for all} \quad t \quad \text{w.p.1} \ . \tag{2.3}$$

With (2.2), the stochastic processes C and S have continuous sample paths. We regard that as the standard case, but we allow C and S to be more general.

With the general framework, the discrete-time fluid-queue model in Section 2.3 is actually a special case of the continuous-time fluid-queue model considered here. The previous discrete-time fluid queue is put in the present framework by letting

$$C(t) \equiv \sum_{k=1}^{\lfloor t \rfloor} V_k \quad \text{and} \quad S(t) \equiv \mu \lfloor t \rfloor, \quad t \geq 0 \ ,$$

where $\lfloor t \rfloor$ is the greatest integer less than or equal to t.

5.2.2. Infinite Capacity

We will consider both the case of unlimited storage space and the case of finite storage space. First suppose that there is unlimited storage space. Let $W(t)$ represent the *workload* (or buffer content, i.e., the quantity of fluid waiting to be processed) at time t. Note that we can have significant fluid flow without ever having any workload. For example, if $W(0) = 0$, $C(t) = \lambda t$ and $S(t) = \mu t$ for all $t \geq 0$, where $\lambda < \mu$, then fluid is processed continuously at rate λ, but $W(t) = 0$ for all t. However, if C is a pure-jump process, then the processing occurs only when $W(t) > 0$. (The workload or virtual-waiting-time process in a standard queue is a pure-jump process.)

The workload $W(t)$ can be defined in terms of an *initial workload* $W(0)$ and a *net-input process* $C(t) - S(t)$, $t \geq 0$, via a *potential-workload process*

$$X(t) \equiv W(0) + C(t) - S(t), \quad t \geq 0 , \tag{2.4}$$

by applying the *one-dimensional reflection map* to X, i.e., by letting

$$W(t) \equiv \phi(X)(t) \equiv X(t) - \inf_{0 \leq s \leq t} \{X(s) \wedge 0\}, \quad t \geq 0 , \tag{2.5}$$

where $a \wedge b = \min\{a, b\}$.

We could incorporate the initial workload $W(0)$ into the cumulative-input process $\{C(t) : t \geq 0\}$ by letting $C(0) = W(0)$. Then X would simply be the net-input process. However, we elect not to do this, because it is convenient to treat the initial conditions separately in the limit theorems.

The potential workload represents what the workload would be if we ignored the emptiness condition, and assumed that there is always output according to the available-processing process S. Then the workload at time t would be $X(t)$: the sum of the initial workload $W(0)$ plus the cumulative input $C(t)$ minus the cumulative output $S(t)$. Since emptiness may sometimes prevent output, we have definition (2.5).

Formula (2.5) is easy to understand by looking at a plot of the potential workload process $\{X(t) : t \geq 0\}$, as shown in Figure 5.1. Figure 5.1 shows a possible sample path of X when $S(t) = \mu t$ for $t \geq 0$ w.p.1 and there is only one on-off source that alternates between busy periods and idle periods, having input rate $r > \mu$ during busy periods and rate 0 during idle periods. Hence the queue alternates between net-input rates $r - \mu > 0$ and $-\mu < 0$. The plot of the potential workload process $\{X(t) : t \geq 0\}$ also can be interpreted as a plot of the actual workload process if we redefine what is meant by the origin. For the workload process, the origin is either 0, if X has not become negative, or the lowest point reached by X. The position of the origin for W is shown by the shaded dashed line in Figure 5.1.

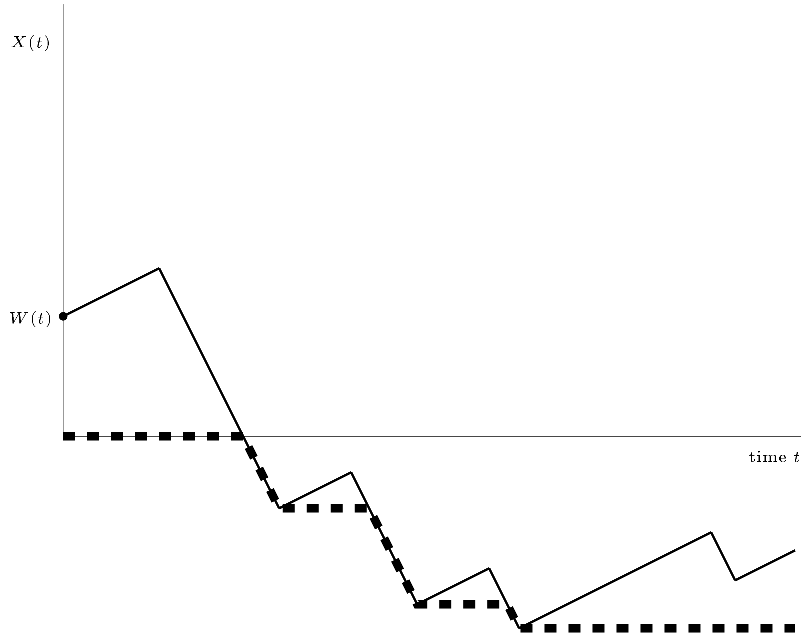

Figure 5.1: A possible realization of the potential workload process $\{X(t) : t \geq 0\}$ and the actual workload process $\{W(t) : t \geq 0\}$ with unlimited storage capacity: The actual workload process appears if the origin is the heavy shaded dashed line; i.e., solid line - dashed line = actual workload.

An important observation is that the single value $W(t)$, for any $t > 0$, depends on the initial segment $\{X(s) : 0 \leq s \leq t\}$. To know $W(t)$, it is not enough to know the single value $X(t)$. However, by (2.5) it is evident that, for any $t > 0$, both $W(t)$ and the initial segment $\{W(s) : 0 \leq s \leq t\}$ are functions of the initial segment $\{X(s) : 0 \leq s \leq t\}$. With appropriate definitions, the reflection map in (2.5) taking the modified net-input process $\{X(t) : t \geq 0\}$ into the workload processes $\{W(t) : t \geq 0\}$ is a continuous function on the space of sample paths; see Section 13.5. Thus, by exploiting the continuous mapping theorem in a function space setting, a limit for a sequence of potential workload processes will translate into a corresponding limit for the associated sequence of workload processes.

Remark 5.2.1. *Model generality.* It may be hard to judge whether the fluid queue model we have introduced is exceptionally general or restrictive. It depends on the perspective: On the one hand, the model is very general because the basic stochastic processes C and S can be almost anything. We illustrate in Chapter 8 by allowing the input C to come from several on-off sources. We are able to treat that more complex model as a special case of the model studied here. On the other hand, the model is also quite restrictive because we assume that the workload stochastic process is directly a reflection of the potential-workload stochastic process. That makes the continuous-mapping approach especially easy to apply. In contrast, as we will see in Chapter 9, it is more difficult to treat the queue-length process in the standard single-server queue without special Markov assumptions. However, additional mathematical analysis shows that the model discrepancy is asymptotically negligible: In the heavy-traffic limit, the queue-length process in the standard single-server queue behaves as if it could be represented directly as a reflection of the associated net-input process. And similar stories hold for other models. The fluid model here is attractive, not only because it is easy to analyze, but also because it captures the essential nature of more complicated models. ■

The general goal in studying this fluid-queue model is to understand how assumed behavior of the basic stochastic processes C and S affects the workload stochastic process W. For example, assuming that the net-input process $C - S$ has stationary increments and negative drift, under minor regularity conditions (see Chapter 1 of Borovkov (1976)), the workload $W(t)$ will have a limiting steady-state distribution. We want to understand how that steady-state distribution depends on the stochastic processes C and S. We also want to describe the transient (time-dependent) behavior of the workload process. Heavy-traffic limits can produce robust approximations that may be useful even when the queue is not in heavy traffic.

We now want to consider the case of a finite storage capacity, but before defining the finite-capacity workload process, we note that the one-sided reflection map in (2.5) can be expressed in an alternative way, which is convenient for treating generalizations such as the finite-capacity model and fluid networks; see Chapter 14 and Harrison (1985) for more discussion. Instead of (2.5), we can write

$$W(t) \equiv \phi(X)(t) \equiv X(t) + L(t)\,, \tag{2.6}$$

where X is the potential workload process in (2.4) and $\{L(t) : t \geq 0\}$ is a nondecreasing "regulator" process that increases only when $W(t) = 0$, i.e., such that

$$\int_0^t W(s)dL(s) = 0, \quad t \geq 0\ . \tag{2.7}$$

From (2.5), we know that

$$L(t) = -\inf_{0 \le s \le t} \{X(s) \wedge 0\}, \quad t \ge 0 . \tag{2.8}$$

It can be shown that the characterization of the reflection map via (2.6) and (2.7) is equivalent to (2.5). For a detailed proof and further discussion, see Chapter 14, which focuses on the more complicated multidimensional generalization.

5.2.3. Finite Capacity

We now modify the definition in (2.6) and (2.7) to construct the finite-capacity workload process. Let the buffer capacity be K. Now we assume that any input that would make the workload process exceed K is lost. Let

$$W(t) \equiv \phi_K(X)(t) \equiv X(t) + L(t) - U(t), \quad t \ge 0 , \tag{2.9}$$

where again $X(t)$ is the potential workload process in (2.4), the initial condition is now assumed to satisfy $0 \le W(0) \le K$, and $L(t)$ and $U(t)$ are both nondecreasing processes. The *lower-boundary regulator process* $L \equiv \psi_L(X)$ increases only when $W(t) = 0$, while the *upper-boundary regulator process* $U \equiv \psi_U(X)$ increases only when $W(t) = K$; i.e., we require that

$$\int_0^t W(s)dL(s) = \int_0^t [K - W(s)]dU(s) = 0, \quad t \ge 0 . \tag{2.10}$$

The random variable $U(t)$ represents the quantity of fluid lost (the overflow) during the interval $[0, t]$. We are often interested in the overflow process $\{U(t) : t \ge 0\}$ as well as the workload process $\{W(t) : t \ge 0\}$.

Note that we can regard the infinite-capacity model as a special case of the finite-capacity model. When $K = \infty$, we can regard the second integral in (2.10) as implying that $U(t) = 0$ for all $t \ge 0$.

Closely paralleling Figure 5.1, for the finite-capacity model we can also depict possible realizations of the processes X and W together, as shown in Figure 5.2. As before, the potential workload process is plotted directly, but we also see the workload (buffer content) process W if we let the origin and upper barrier move according to the two heavily shaded dashed lines, which remain a distance K apart. Decreases in the dashed lines correspond to increases in the lower-barrier regulator process L, while increases in the shaded lines correspond to increases in the upper-barrier regulator process U. From the Figure 5.2, the validity of (2.9) and (2.10) is evident. Furthermore, it is evident that the two-sided reflection in (2.9) can be defined by successive applications of the one-sided reflection map in (2.5) and (2.6) corresponding to the lower and upper barriers separately. For further discussion, see Section 14.8.

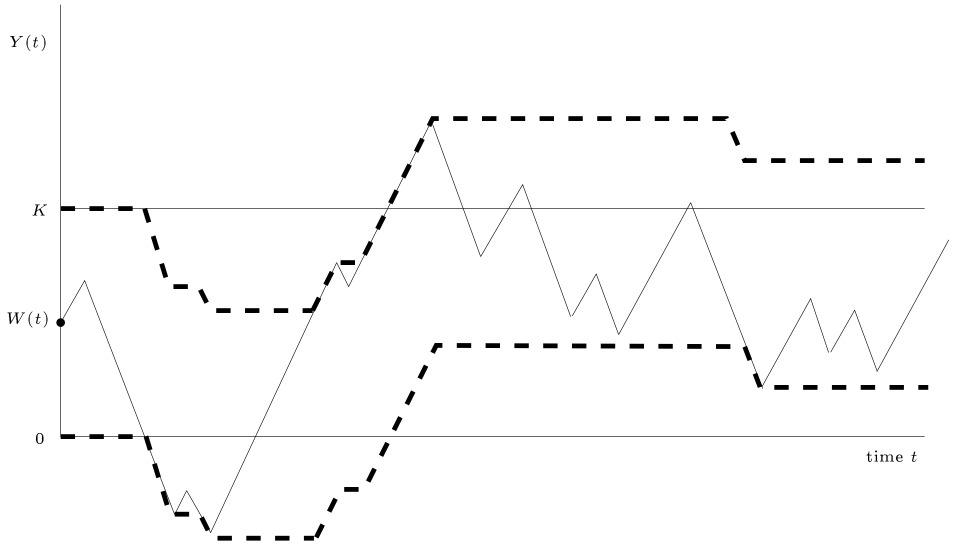

Figure 5.2: A possible realization of the potential workload process $\{X(t) : t \geq 0\}$ and the actual workload process $\{W(t) : t \geq 0\}$ with finite storage capacity K: The actual workload process appears if the origin and upper limit are the heavily shaded dashed lines always a distance K apart. As in Figure 5.1, solid line - lower dashed line = actual workload.

As in the infinite-capacity case, given K, the initial segment $\{W(s), L(s), U(s) : 0 \leq s \leq t\}$ depends on the potential-workload process X via the corresponding initial segment $\{X(s) : 0 \leq s \leq t\}$. Again, under regularity conditions, the reflection map in (2.9) taking $\{X(t) : t \geq 0\}$ into $\{(W(t), L(t), U(t) : t \geq 0\}$ is a continuous function on the space of sample paths (mapping initial segments into initial segments). Thus, stochastic-process limits for X translate into stochastic-process limits for (W, L, U), by exploiting the continuous-mapping approach with the full reflection map (ϕ_K, ψ_L, ψ_U) in a function space setting.

Let $D(t)$ represent the amount of fluid processed (not counting any overflow) during the time interval $[0, t]$. We call $\{D(t) : t \geq 0\}$ the *departure process*.

From (2.4) and (2.9),

$$\begin{aligned} D(t) &= W(0) + C(t) - W(t) - U(t) \\ &= S(t) - L(t), \quad t \geq 0 \ . \end{aligned} \tag{2.11}$$

Note that the departure process D in (2.11) is somewhat more complicated than the workload process W because, unlike the workload process, the departure process cannot be represented directly as a function of the potential workload process X or the net-input process $C - S$. In general, the departure process cannot be represented directly in terms of X or $C - S$ because these processes cannot see the values of jumps in C and S that occur at the same time. Simultaneous jumps in C and S correspond to instants at which fluid arrives and some of it is instantaneously processed. The fluid that is instantaneously processed immediately upon arrival never affects the workload process. To obtain stochastic-process limits for the departure process, we will impose a condition to rule out such cancelling jumps in the limit processes associated with C and S. In particular, the departure process is considerably less complicated in the case of constant processing, as in (2.1).

We may also be interested in the *processing time* $T(t)$, i.e., the time required to process the work in the system at any time t, not counting any future input. For the processing time to correctly represent the actual processing time for the last particle of fluid in the queue, the fluid must be processed in the order of arrival. The processing time $T(t)$ is the first passage time to the level $W(t)$ by the future-available-processing process $\{S(t+u) - S(t) : u \geq 0\}$, i.e.,

$$T(t) \equiv inf\{u \geq 0 : S(t+u) - S(t) \geq W(t)\}, \quad t \geq 0 \ . \tag{2.12}$$

We can obtain an equivalent representation, involving a first passage time of the process S alone on the left in the infimum, if we use formula (2.9) for $W(t)$:

$$\begin{aligned} T(t) + t &= t + inf\{u \geq 0 : S(t+u) - S(t) \geq X(t) + L(t) - U(t)\}, \\ &= inf\{u \geq 0 : S(u) \geq W(0) + C(t) + L(t) - U(t)\}, \quad t \geq 0 \ . \end{aligned} \tag{2.13}$$

In general, the processing time is relatively complicated, but in the common case of constant processing in (2.1), $T(t)$ is a simple modification of $W(t)$, namely,

$$T(t) = W(t)/\mu, \quad t \geq 0 \ . \tag{2.14}$$

More generally, heavy-traffic limits also lead to such simplifications; see Section 5.9.2.

5.3. Unstable Queues

There are two main reasons queues experience congestion (which here means buildup of workload): First, the queue may be *unstable* (or overloaded); i.e., the input rate may exceed the output rate for an extended period of time, when there is ample storage capacity. Second, the queue may be stable, i.e., the long-run input rate may be less than the long-run output rate, but nevertheless short-run fluctuations produce temporary periods during which the input exceeds the output.

The unstable case tends to produce more severe congestion, but the stable case is more common, because systems are usually designed to be stable. Unstable queues typically arise in the presence of system failures. Since there is interest in system performance in the presence of failures, there is interest in the performance of unstable queues. For our discussion of unstable queues, we assume that there is unlimited storage capacity. We are interested in the buildup of congestion, which is described by the transient (or time-dependent) behavior of the queueing processes.

5.3.1. Fluid Limits for Fluid Queues

For unstable queues, useful insight can be gained from *fluid limits* associated with functional laws of large numbers (FLLN's). These stochastic-process limits are called fluid limits because the limit processes are deterministic functions of the form ct for some constant c. (More generally, with time-varying input and output rates, the limits could be deterministic functions of the form $\int_0^t r(s)ds$, $t \geq 0$, for some deterministic integrable function r.)

To express the FLLN's, we scale space and time both by n. As before, we use bold capitals to represent the scaled stochastic processes and associated limiting stochastic processes in the function space D. We use a hat to denote scaled stochastic processes with the fluid scaling (scaling space as well as time by n). Given the stochastic processes defined for the fluid-queue model in the previous section, form the associated scaled stochastic processes

$$\begin{aligned}
\hat{\mathbf{C}}_n(t) &\equiv n^{-1}C(nt), \\
\hat{\mathbf{S}}_n(t) &\equiv n^{-1}S(nt), \\
\hat{\mathbf{X}}_n(t) &\equiv n^{-1}X(nt), \\
\hat{\mathbf{W}}_n(t) &\equiv n^{-1}W(nt), \\
\hat{\mathbf{L}}_n(t) &\equiv n^{-1}L(nt), \\
\hat{\mathbf{D}}_n(t) &\equiv n^{-1}D(nt),
\end{aligned}$$

$$\hat{\mathbf{T}}_n(t) \equiv n^{-1}T(nt), \quad t \geq 0 . \qquad (3.1)$$

The continuous-mapping approach shows that FLLN's for C and S imply a joint FLLN for all the processes. As before, let $\mathbf{e}$ be the identity map, i.e., $\mathbf{e}(t) = t$, $t \geq 0$. Let $\mu \wedge \lambda \equiv min\{\mu, \lambda\}$ and $\lambda^+ \equiv max\{\lambda, 0\}$ for constants λ and μ.

We understand D to be the space $D([0,\infty),\mathbb{R})$, endowed with either the J_1 or the M_1 topology, as defined in Section 3.3. Since the limits are continuous deterministic functions, the J_1 and M_1 topologies here are equivalent to uniform convergence on compact subintervals. As in Section 3.3, we use D^k to denote the k-dimensional product space with the product topology; then $x_n \to x$, where $x_n \equiv (x_n^1, \ldots x_n^k)$ and $x \equiv (x^1, \ldots, x^k)$, if and only if $x_n^i \to x_i$ for each i.

We first establish a functional weak law of large numbers (FWLLN), involving convergence in probability or, equivalently (because of the deterministic limit), convergence in distribution (see p. 27 of Billingsley (1999)). As indicated above, we restrict attention to the infinite-capacity model. It is easy to extend the results to the finite-capacity model, provided that the capacity is allowed to increase with n, as in Section 2.3.

Theorem 5.3.1. (FWLLN for the fluid queue) *In the infinite-capacity fluid-queue model, if* $\hat{\mathbf{C}}_n \Rightarrow \lambda\mathbf{e}$ *and* $\hat{\mathbf{S}}_n \Rightarrow \mu\mathbf{e}$ *in* (D, M_1)*, where* $0 < \mu < \infty$ *and* $\hat{\mathbf{C}}_n$ *and* $\hat{\mathbf{S}}_n$ *are given in* (3.1)*, then*

$$\begin{aligned}&(\hat{\mathbf{C}}_n, \hat{\mathbf{S}}_n, \hat{\mathbf{X}}_n, \hat{\mathbf{W}}_n, \hat{\mathbf{L}}_n, \hat{\mathbf{D}}_n, \hat{\mathbf{T}}_n) \Rightarrow \\ &\quad (\lambda\mathbf{e}, \mu\mathbf{e}, (\lambda-\mu)\mathbf{e}, (\lambda-\mu)^+\mathbf{e}, (\mu-\lambda)^+\mathbf{e}, (\lambda\wedge\mu)\mathbf{e}, (\rho-1)^+\mathbf{e}) \qquad (3.2)\end{aligned}$$

in $(D, M_1)^7$ *for* $\rho \equiv \lambda/\mu$.

Proof. The single limits can be combined into joint limits because the limits are deterministic, by virtue of Theorem 11.4.5. So start with the joint convergence

$$(\hat{\mathbf{C}}_n, \hat{\mathbf{S}}_n, n^{-1}W(0)) \Rightarrow (\lambda\mathbf{e}, \mu\mathbf{e}, 0) \quad \text{in} \quad (D, M_1)^2 \times \mathbb{R} .$$

Since

$$\hat{\mathbf{X}}_n = \hat{\mathbf{C}}_n - \hat{\mathbf{S}}_n + n^{-1}W(0)$$

by (2.4), we can apply the continuous-mapping approach with addition, using the fact that addition on D^2 is measurable and continuous almost surely with respect to the limit process, to get the limit

$$\hat{\mathbf{X}}_n \Rightarrow \hat{\mathbf{X}} \equiv (\lambda - \mu)\mathbf{e} .$$

Specifically, we invoke Theorems 3.4.3 and 12.7.3 and Remark 12.7.1.

Then, because of (2.5) – (2.8), we can apply the simple continuous-mapping theorem, Theorem 3.4.1, with the reflection map to get

$$\hat{\mathbf{W}}_n \Rightarrow \hat{\mathbf{W}} \equiv \phi(\hat{\mathbf{X}}) = (\lambda - \mu)^+ \mathbf{e}$$

and

$$\hat{\mathbf{L}}_n \Rightarrow \hat{\mathbf{L}} \equiv \psi_L(\hat{\mathbf{X}}) = (\mu - \lambda)^+ \mathbf{e} \ ,$$

drawing on Theorems 13.5.1, 13.4.1 and 14.8.5. Then, by (2.11), we can apply the continuous-mapping approach with addition again to obtain $\hat{\mathbf{D}}_n \Rightarrow \hat{\mathbf{D}} = (\lambda \wedge \mu)\mathbf{e}$. Finally, by (2.13),

$$n^{-1}T(nt) + t = inf\{u \geq 0 : n^{-1}S(nu) \geq n^{-1}(C(nt) + L(nt) + W(0))\} \quad (3.3)$$

or, in more compact notation,

$$\hat{\mathbf{T}}_n + \mathbf{e} = \hat{\mathbf{S}}_n^{-1} \circ (\hat{\mathbf{C}}_n + \hat{\mathbf{L}}_n + n^{-1}W(0)) \ . \quad (3.4)$$

Hence, we can again apply the continuous-mapping approach, this time with the inverse and composition functions. As with addition used above, these functions as maps from D and $D \times D$ to D are measurable and continuous almost surely with respect to the deterministic, continuous, strictly increasing limits. Specifically, by Corollary 13.6.4 and Theorem 13.2.1, we obtain

$$\hat{\mathbf{T}}_n + \mathbf{e} \Rightarrow \mu^{-1}\mathbf{e} \circ (\lambda \mathbf{e} + (\mu - \lambda)^+ \mathbf{e}) = (\rho \vee 1)\mathbf{e} \ ,$$

so that

$$\hat{\mathbf{T}}_n \Rightarrow (\rho - 1)^+ \mathbf{e} \ ,$$

as claimed. By Theorem 11.4.5, all limits can be joint. ■

From Theorem 5.3.1, we can characterize stable queues and unstable queues by the conditions $\lambda \leq \mu$ and $\lambda > \mu$, respectively, where λ and μ are the translation constants in the limits for the input process C and the available-processing process S. Equivalently, we can use the *traffic intensity* ρ, defined as

$$\rho \equiv \lambda / \mu \ . \quad (3.5)$$

From the relatively crude fluid-limit perspective, there is no congestion if $\rho \leq 1$; i.e., Theorem 5.3.1 implies that $\hat{\mathbf{W}}_n \Rightarrow 0\mathbf{e}$ if $\rho \leq 1$. On the other hand, if $\rho > 1$, then the workload tends to grow linearly at rate $\lambda - \mu$. Consistent with intuition, the fluid limits suggest using a simple deterministic analysis to describe congestion in unstable queues. When a queue is unstable for a significant time, the relatively simple deterministic analysis may capture the dominant congestion effect. The same reasoning applies to queues with time-dependent input and output rates that are unstable for substantial periods of time. See Oliver and Samuel (1962), Newell (1982) and Hall (1991) for discussions of direct deterministic analysis of the congestion in queues.

Ordinary weak laws of large numbers (WLLN's), such as

$$t^{-1}W(t) \Rightarrow (\lambda - \mu)^{+} \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad t \to \infty ,$$

follow immediately from the FWLLN's in Theorem 5.3.1 by applying the continuous-mapping approach with the projection map, which maps a function x into $x(1)$. We could not obtain these WLLN's or the stronger FWLLN's in Theorem 5.3.1 if we assumed only ordinary WLLN's for C and S, i.e., if we had started with limits such as

$$t^{-1}C(t) \Rightarrow \lambda \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad t \to \infty ,$$

because we needed to exploit the continuous-mapping approach in the function space D. We cannot go directly from a WLLN to a FWLLN, because a FWLLN is strictly stronger than a WLLN.

However, we can obtain functional strong laws of large numbers (FSLLN's) starting from ordinary strong laws of large numbers (SLLN's), because a SLLN implies a corresponding FSLLN; see Theorem 3.2.1 and Corollary 3.2.1 in the Internet Supplement. To emphasize that point, we now state the SLLN version of Theorem 5.3.1. Once we go from the SLLN's for C and S to the FSLLN's, the proof is the same as for Theorem 5.3.1.

Theorem 5.3.2. (FSLLN for the fluid queue) *In the infinite-capacity fluid-queue model, if*

$$t^{-1}C(t) \to \lambda \quad \textit{and} \quad t^{-1}S(t) \to \mu \quad \textit{in} \quad \mathbb{R} \quad \textit{w.p.1} \quad \textit{as} \quad t \to \infty ,$$

for $0 < \mu < \infty$, *then*

$$(\hat{\mathbf{C}}_n, \hat{\mathbf{S}}_n, \hat{\mathbf{X}}_n, \hat{\mathbf{W}}_n, \hat{\mathbf{L}}_n, \hat{\mathbf{D}}_n, \hat{\mathbf{T}}_n) \to \\ (\lambda\mathbf{e}, \mu\mathbf{e}, (\lambda - \mu)\mathbf{e}, (\lambda - \mu)^{+}\mathbf{e}, (\mu - \lambda)^{+}\mathbf{e}, (\lambda \wedge \mu)\mathbf{e}, (\rho - 1)^{+}\mathbf{e}) \quad (3.6)$$

w.p.1 in $(D, M_1)^7$ *for* ρ *in (3.5).*

5.3.2. Stochastic Refinements

We can also employ stochastic-process limits to obtain a more detailed description of congestion in unstable queues. These stochastic-process limits yield *stochastic refinements to the fluid limits* in Theorems 5.3.1 and 5.3.2 above. For the stochastic refinements, we introduce new scaled stochastic processes:

$$\begin{aligned}
\mathbf{C}_n(t) &\equiv c_n^{-1}(C(nt) - \lambda nt), \\
\mathbf{S}_n(t) &\equiv c_n^{-1}(S(nt) - \mu nt), \\
\mathbf{X}_n(t) &\equiv c_n^{-1}(X(nt) - (\lambda - \mu)nt), \\
\mathbf{W}_n(t) &\equiv c_n^{-1}(W(nt) - (\lambda - \mu)^+ nt), \\
\mathbf{L}_n(t) &\equiv c_n^{-1}(L(nt) - (\mu - \lambda)^+ nt), \\
\mathbf{D}_n(t) &\equiv c_n^{-1}(D(nt) - (\lambda \wedge \mu)nt), \\
\mathbf{T}_n(t) &\equiv c_n^{-1}(T(nt) - (\rho - 1)^+ nt), \quad t \geq 0 .
\end{aligned} \tag{3.7}$$

As in the last chapter, the space scaling constants will be assumed to satisfy $c_n \to \infty$ and $n/c_n \to \infty$ as $n \to \infty$. The space-scaling constants will usually be a power, i.e., $c_n = n^H$ for $0 < H < 1$, but we allow other possibilities. In the following theorem we only discuss the cases $\rho < 1$ and $\rho > 1$. The more complex boundary case $\rho = 1$ is covered as a special case of results in the next section. Recall that D^k is the product space with the product topology; here we let the component space $D \equiv D^1$ have either the J_1 or the M_1 topology.

Since the limit processes $\mathbf{C}$ and $\mathbf{S}$ below may now have discontinuous sample paths, we need an extra condition to apply the continuous-mapping approach with addition. The extra condition depends on random sets of discontinuity points; e.g.,

$$Disc(\mathbf{S}) \equiv \{t : \mathbf{S}(t) \neq \mathbf{S}(t-)\} ,$$

where $x(t-)$ is the left limit of the function x in D (see Section 12.2). The random *set of common discontinuity points* of $\mathbf{C}$ and $\mathbf{S}$ is $Disc(\mathbf{C}) \cap Disc(\mathbf{S})$. The *jump* in $\mathbf{S}$ associated with a discontinuity at t is $\mathbf{S}(t) - \mathbf{S}(t-)$. The required extra condition is somewhat weaker for the M_1 topology than for the J_1 topology.

Theorem 5.3.3. (FCLT's for the stable and unstable fluid queues) *In the infinite-capacity fluid queue, suppose that $c_n \to \infty$ and $c_n/n \to 0$ as $n \to \infty$.*

Suppose that

$$(\mathbf{C}_n, \mathbf{S}_n) \Rightarrow (\mathbf{C}, \mathbf{S}) \quad in \quad D^2 , \tag{3.8}$$

where D^2 has the product topology with the topology on D^1 being either J_1 or M_1, $\mathbf{C}_n$ and $\mathbf{S}_n$ are defined in (3.7) *and*

$$P(\mathbf{C}(0) = \mathbf{S}(0) = 0) = 1 . \tag{3.9}$$

If the topology is J_1, assume that $\mathbf{C}$ and $\mathbf{S}$ almost surely have no common discontinuities. If the topology is M_1, assume that $\mathbf{C}$ and $\mathbf{S}$ almost surely have no common discontinuities with jumps of common sign.

(a) If $\rho < 1$ and $\mathbf{C} - \mathbf{S}$ has no positive jumps, then

$$(\mathbf{C}_n, \mathbf{S}_n, \mathbf{X}_n, \mathbf{W}_n, \mathbf{L}_n, \mathbf{D}_n) \Rightarrow (\mathbf{C}, \mathbf{S}, \mathbf{C} - \mathbf{S}, 0\mathbf{e}, \mathbf{S} - \mathbf{C}, \mathbf{C}) \tag{3.10}$$

in D^6 with the same topology.

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

$$(\mathbf{C}_n, \mathbf{S}_n, \mathbf{X}_n, \mathbf{W}_n, \mathbf{L}_n, \mathbf{D}_n) \Rightarrow (\mathbf{C}, \mathbf{S}, \mathbf{C} - \mathbf{S}, \mathbf{C} - \mathbf{S}, 0\mathbf{e}, \mathbf{S}) \tag{3.11}$$

in D^6 with the same topology.

Proof. Paralleling the proof of Theorem 5.3.1 above, we start by applying condition (3.8) and Theorem 11.4.5 to obtain the joint convergence

$$(\mathbf{C}_n, \mathbf{S}_n, c_n^{-1} W(0)) \Rightarrow (\mathbf{C}, \mathbf{S}, 0) \quad \text{in} \quad D^2 \times \mathbb{R} .$$

Then, as before, we apply the continuous mapping approach with addition, now invoking the conditions on the discontinuities of $\mathbf{C}$ and $\mathbf{S}$, to get

$$(\mathbf{C}_n, \mathbf{S}_n, \mathbf{X}_n, c_n^{-1} W(0)) \Rightarrow (\mathbf{C}, \mathbf{S}, \mathbf{C} - \mathbf{S}, 0) \quad \text{in} \quad D^3 \times \mathbb{R} . \tag{3.12}$$

For the M_1 topology, we apply Theorems 3.4.3 and 12.7.3 and Remark 12.7.1. For J_1, we apply the J_1 analog of Corollary 12.7.1; see Remark 12.6.2.

The critical step is treating $\mathbf{W}_n$. For that purpose, we apply Theorem 13.5.2, for which we need to impose the extra condition that $C - S$ have no positive jumps in part (a). We also use condition (3.9), but it can be weakened. We can use the Skorohod representation theorem, Theorem 3.2.2, to carry out the argument for individual sample paths.

The limit for $\mathbf{L}_n$ in part (a) then follows from (2.6), again exploiting the continuous-mapping approach with addition. The limits for $\mathbf{L}_n$ in part

(b) follows from Theorem 13.4.4, using (2.8) and condition (3.9). We can apply the convergence-together theorem, Theorem 11.4.7, to get limits for the scaled departure process $\mathbf{D}_n$. If $\lambda < \mu$, then

$$d_t(\mathbf{D}_n, \mathbf{C}_n) \leq \|\mathbf{D}_n - \mathbf{C}_n\|_t \leq \|c_n^{-1}W(0) - \mathbf{W}_n\|_t \Rightarrow 0$$

by (2.11), where d_t and $\|\cdot\|$ are the J_1 (or M_1) and uniform metrics for the time interval $[0,t]$, as in equations (3.2) and (3.1) of Section 3.3. If $\lambda > \mu$, then

$$d_t(\mathbf{D}_n, \mathbf{S}_n) \leq \|\mathbf{D}_n - \mathbf{S}_n\|_t \leq \|\mathbf{L}_n\|_t \Rightarrow 0$$

by (2.11). ∎

The obvious sufficient condition for the limit processes $\mathbf{C}$ and $\mathbf{S}$ to almost surely have no discontinuities with jumps of common sign is to have no common discontinuities at all. For that, it suffices for $\mathbf{C}$ and $\mathbf{S}$ to be independent processes without any fixed discontinuities; i.e., $\mathbf{C}$ has no fixed discontinuities if $P(t \in Disc(\mathbf{C})) = 0$ for all t.

With the J_1 topology, the conclusion can be strengthened to the strong SJ_1 topology instead of the product J_1 topology, but that is not true for M_1; see Remark 9.3.1 and Example 14.5.1.

When $\rho < 1$, we not only obtain the zero fluid limit $\hat{\mathbf{W}}_n \Rightarrow 0\mathbf{e}$ in Theorem 5.3.1, but we also obtain the zero limit $\mathbf{W}_n \Rightarrow 0\mathbf{e}$ in Theorem 5.3.3 (a) with the refined scaling in (3.7), provided that $C - S$ has no positive jumps. However, if $C - S$ has positive jumps, then the scaled workload process $\mathbf{W}_n$ fails to be uniformly negligible. That shows the impact of jumps in the limit process.

Under extra conditions, we get a limit for $\mathbf{T}_n$ jointly with the limit in Theorem 5.3.3.

Theorem 5.3.4. (FCLT for the processing time) *Let the conditions of Theorem* 5.3.3 *hold. If the topology is J_1, assume that S has no positive jumps.*

(a) If $\rho < 1$, then jointly with the limit in (3.10)

$$\mathbf{T}_n \Rightarrow 0\mathbf{e}$$

in D with the same topology.

(b) Suppose that $\rho > 1$. If the topology is J_1, assume that $\mathbf{C}$ and $\mathbf{S} \circ \rho\mathbf{e}$ almost surely have no common discontinuities. If the topology is M_1, assume that $\mathbf{C}$ and $\mathbf{S} \circ \rho\mathbf{e}$ almost surely have no common discontinuities with jumps of common sign. Then jointly with the limit in (3.11)

$$\mathbf{T}_n \Rightarrow \mu^{-1}(\mathbf{C} - \mathbf{S} \circ \rho\mathbf{e})$$

in D with the same topology.

Proof. We can apply Theorem 13.7.4 to treat $\mathbf{T}_n$, starting from (3.3) and (3.4). If $\lambda > \mu$, then

$$(n/c_n)(\hat{\mathbf{S}}_n - \mu\mathbf{e}, \hat{\mathbf{C}}_n + \hat{\mathbf{L}}_n + n^{-1}W(0) - \lambda\mathbf{e}) \Rightarrow (\mathbf{S}, \mathbf{C}) \ , \tag{3.13}$$

because $\mathbf{L}_n \Rightarrow 0\mathbf{e}$ and $n^{-1}W(0) \Rightarrow 0$. If $\lambda < \mu$, then

$$(n/c_n)(\hat{\mathbf{S}}_n - \mu\mathbf{e}, \hat{\mathbf{C}}_n + \hat{\mathbf{L}}_n + n^{-1}W(0) - \mu\mathbf{e}) \Rightarrow (\mathbf{S}, \mathbf{S}) \ , \tag{3.14}$$

because, by (2.6),

$$d_t(\mathbf{C}_n + \mathbf{L}_n + c_n^{-1}W(0), \mathbf{S}_n) \le \|\mathbf{L}_n + \mathbf{X}_n\|_t = \|\mathbf{W}_n\|_t \Rightarrow 0 \ .$$

We can apply Theorem 13.7.4 to obtain limits for $\mathbf{T}_n$ jointly with the other limits because

$$\begin{aligned} \mathbf{T}_n &= (n/c_n)(\hat{\mathbf{T}}_n - (\rho - 1)^+\mathbf{e}) \\ &= (n/c_n)(\hat{\mathbf{S}}_n^{-1} \circ \hat{\mathbf{Z}}_n - (\rho \vee 1)\mathbf{e}) \\ &= (n/c_n)(\hat{\mathbf{S}}_n^{-1} \circ \hat{\mathbf{Z}}_n - \mu^{-1}\mathbf{e} \circ (\lambda \vee \mu)\mathbf{e}) \end{aligned}$$

for appropriate $\mathbf{Z}_n$ (specified in (3.13) and (3.14) above), where $n/c_n \to \infty$ as $n \to \infty$. Theorem 13.7.4 requires condition (3.9) for $\mathbf{S}$. ■

We regard the unstable case $\rho > 1$ as the case of primary interest for a single model. When $\rho > 1$, Theorem 5.3.3 (b) concludes that $W(t)$ obeys the same FCLT as $X(t)$. In a long time scale, the amount of reflection is negligible. Thus we obtain the approximation

$$W(t) \approx (\lambda - \mu)t + c_n\mathbf{X}(t/n) \tag{3.15}$$

for the workload, where $\mathbf{X} = \mathbf{C} - \mathbf{S}$. In the common setting of Donsker's theorem, $c_n = n^{1/2}$ and $\mathbf{X} = \sigma_X\mathbf{B}$, where $\mathbf{B}$ is standard Brownian motion. In that special case, (3.15) becomes

$$\begin{aligned} W(t) &\approx (\lambda - \mu)t + n^{1/2}\sigma_X\mathbf{B}(t/n) \\ &\approx N((\lambda - \mu)t, \sigma_X^2 t) \ . \end{aligned} \tag{3.16}$$

In this common special case, the stochastic refinement of the LLN shows that the workload obeys a CLT and, thus, the workload $W(t)$ should be approximately normally distributed with mean equal to the fluid limit $(\lambda - \mu)t$ and standard deviation proportional to $\sqrt{t}$, with the variability parameter given explicitly. With heavy tails or strong dependence (or both), but still with finite mean, the stochastic fluctuations about the mean will be greater, as is made precise by the stochastic-process limits.

Remark 5.3.1. *Implications for queues in series.* Part (a) of Theorem 5.3.3 has important implications for queues in series: If the first of two queues is stable with $\rho < 1$, then the departure process D at the first queue obeys the same FCLT as the input process C at that first queue. Thus, if we consider a heavy-traffic limit for the second queue (either because the second queue is unstable or because we consider a sequence of models for the second queue with the associated sequence of traffic intensities at the second queue approaching the critical level for stability, as in the next section), then the heavy-traffic limit at the second queue depends on the first queue only through the input stochastic process at that first queue. In other words, the heavy-traffic behavior of the second queue is the same as if the first queue were not even there. We obtain more general and more complicated heavy-traffic stochastic-process limits for the second queue only if we consider a sequence of models for both queues, and simultaneously let the sequences of traffic intensities at both queues approach the critical levels for stability, which puts us in the setting of Chapter 14. For further discussion, see Example 9.9.1, Chapter 14 and Karpelovich and Kreinin (1994). ∎

In this section we have seen how heavy-traffic stochastic-process limits can describe the congestion in an unstable queue. We have considered the relatively elementary case of constant input and output rates. Variations of the same approach apply to queues with time-varying input and output rates; see Massey and Whitt (1994a), Mandelbaum and Massey (1995), Mandelbaum, Massey and Reiman (1998) and Chapter 9 of the Internet Supplement.

5.4. Heavy-Traffic Limits for Stable Queues

We now want to establish nondegenerate heavy-traffic stochastic-process limits for stochastic processes in stable fluid queues (where the long-run input rate is less than the maximum potential output rate). (With a finite storage capacity, the workload will of course remain bounded even if the long-run input rate exceeds the output rate.)

The first heavy-traffic limits for queues were established by Kingman (1961, 1962, 1965). The treatment here is in the spirit of Iglehart and Whitt (1970a, b) and Whitt (1971a), although those papers focused on standard queueing models, as considered here in Chapters 9 and 10. An early heavy-traffic limit for finite-capacity queues was established by Kennedy

(1973). See Whitt (1974b) and Borovkov (1976, 1984) for background on early heavy-traffic limits.

In order to establish the heavy-traffic stochastic-process limits for stable queues, we consider a sequence of models indexed by a subscript n, where the associated sequence of traffic intensities $\{\rho_n : n \geq 1\}$ converges to 1, the critical level for stability, as $n \to \infty$. We have in mind the case in which the traffic intensities approach 1 from below, denoted by $\rho_n \uparrow 1$, but that is not strictly required. For each n, there is a cumulative-input process C_n, an available-processing process S_n, a storage capacity K_n with $0 < K_n \leq \infty$ and an initial workload $W_n(0)$ satisfying $0 \leq W_n(0) \leq K_n$. As before, we make no specific structural or stochastic assumptions about the stochastic processes C_n and S_n, so we have very general models. A more detailed model for the input is considered in Chapter 8.

To have the traffic intensity well defined in our setting, we assume that the limits

$$\lambda_n \equiv \lim_{t \to \infty} t^{-1} C_n(t) \tag{4.1}$$

and

$$\mu_n \equiv \lim_{t \to \infty} t^{-1} S_n(t) \tag{4.2}$$

exist w.p.1 for each n. We call λ_n the *input rate* and μ_n the *maximum potential output rate* for model n. (The actual output rate is the input rate minus the overflow rate.) Then the traffic intensity in model n is

$$\rho_n \equiv \lambda_n / \mu_n \ . \tag{4.3}$$

We will be letting $\rho_n \to 1$ as $n \to \infty$.

Given the basic model elements above, we can construct the potential-workload processes $\{X_n(t) : t \geq 0\}$, the workload processes $\{W_n(t) : t \geq 0\}$, the upper-barrier regulator (overflow) processes $\{U_n(t) : t \geq 0\}$, the lower-barrier regulator processes $\{L_n(t) : t \geq 0\}$ and the departure processes $\{D_n(t) : t \geq 0\}$ as described in Sections 5.2.

We now form associated scaled processes. We could obtain fluid limits in this setting, paralleling Theorems 5.3.1 and 5.3.2, but they add little beyond the previous results. Hence we go directly to the generalizations of Theorem 5.3.3. We scale the processes as in (3.7), but now we have processes and translation constants for each n. Let

$$\begin{aligned}
\mathbf{C}_n(t) &\equiv c_n^{-1}(C_n(nt) - \lambda_n nt) \ , \\
\mathbf{S}_n(t) &\equiv c_n^{-1}(S_n(nt) - \mu_n nt) \ , \\
\mathbf{X}_n(t) &\equiv c_n^{-1} X_n(nt) \ ,
\end{aligned}$$

$$\begin{aligned}
\mathbf{W}_n(t) &\equiv c_n^{-1} W_n(nt) \; , \\
\mathbf{U}_n(t) &\equiv c_n^{-1} U_n(nt) \; , \\
\mathbf{L}_n(t) &\equiv c_n^{-1} L_n(nt) \; , \quad t \ge 0 \; .
\end{aligned} \tag{4.4}$$

For the scaling constants, we have in mind $\lambda_n \to \lambda$ and $\mu_n \to \mu$ as $n \to \infty$, where $0 < \lambda < \infty$ and $0 < \mu < \infty$, with $c_n \to \infty$ and $n/c_n \to \infty$ as $n \to \infty$. As in Section 2.3, the upper barrier must grow as $n \to \infty$; specifically, we require that $K_n = c_n K$.

Our key assumption is a joint limit for $\mathbf{C}_n$ and $\mathbf{S}_n$ in (4.4). When there are limits for $\mathbf{C}_n$ and $\mathbf{S}_n$ with the translation terms involving λ_n and μ_n, the w.p.1 limits in (4.1) and (4.2) usually hold too, but (4.1) and (4.2) are actually not required. However, convergence in probability in (4.1) and (4.2) follows directly as a consequence of the convergence in distribution assumed below. Hence it is natural for the limits in (4.1) and (4.2) to hold as well.

Let (ϕ_K, ψ_U, ψ_L) be the reflection map mapping a potential-workload process X into the triple (W, U, L), as defined in Section 5.2. Here is the general heavy-traffic stochastic-process limit for stable fluid queues. It follows directly from the continuous-mapping approach using addition and reflection.

Theorem 5.4.1. (general heavy-traffic limit for stable fluid queues) *Consider a sequence of fluid queues indexed by n with capacities K_n, $0 < K_n \le \infty$, general cumulative-input processes $\{C_n(t) : t \ge 0\}$ and general cumulative-available-processing processes $\{S_n(t) : t \ge 0\}$. Suppose that $K_n = c_n K$, $0 < K \le \infty$, $0 \le W_n(0) \le K_n$,*

$$(c_n^{-1} W_n(0), \mathbf{C}_n, \mathbf{S}_n) \Rightarrow (W'(0), \mathbf{C}, \mathbf{S}) \quad \text{in} \quad \mathbb{R} \times D^2 \tag{4.5}$$

for $\mathbf{C}_n$ and $\mathbf{S}_n$ in (4.4), where the topology on D^2 is the product topology with the topology on D^1 being either J_1 or M_1, $c_n \to \infty$, $c_n/n \to 0$ and $\lambda_n - \mu_n \to 0$, so that

$$\eta_n \equiv n(\lambda_n - \mu_n)/c_n \to \eta \; , \tag{4.6}$$

where $-\infty < \eta < \infty$. If the topology is J_1, suppose that almost surely $\mathbf{C}$ and $\mathbf{S}$ have no common discontinuities. If the topology is M_1, suppose that almost surely $\mathbf{C}$ and $\mathbf{S}$ have no common discontinuities with jumps of common sign. Then, jointly with the limit in (4.5),

$$(\mathbf{X}_n, \mathbf{W}_n, \mathbf{U}_n, \mathbf{L}_n) \Rightarrow (\mathbf{X}, \mathbf{W}, \mathbf{U}, \mathbf{L}) \tag{4.7}$$

in D^4 with the same topology, where

$$\mathbf{X}(t) = W'(0) + \mathbf{C}(t) - \mathbf{S}(t) + \eta t, \quad t \ge 0 \ . \tag{4.8}$$

and

$$(\mathbf{W}, \mathbf{U}, \mathbf{L}) \equiv (\phi_K(\mathbf{X}), \psi_U(\mathbf{X}), \psi_L(\mathbf{X})) \tag{4.9}$$

with (ϕ_K, ψ_U, ψ_L) being the reflection map associated with capacity K.

Proof. Note that

$$\mathbf{X}_n = c_n^{-1} W_n(0) + \mathbf{C}_n - \mathbf{S}_n + \eta_n \mathbf{e} \ , \tag{4.10}$$

where $\mathbf{e}(t) \equiv t$ for $t \ge 0$. Thus, just as in Theorems 5.3.1 and 5.3.3 above, we can apply the continuous-mapping approach starting from the joint convergence

$$(c_n^{-1} W_n(0), \mathbf{C}_n, \mathbf{S}_n, \eta_n \mathbf{e}) \Rightarrow (W'(0), \mathbf{C}, \mathbf{S}, \eta \mathbf{e}) \tag{4.11}$$

in $\mathbb{R} \times D^3$, which follows from (4.5), (4.6) and Theorem 11.4.5. We apply the continuous mapping theorem, Theorem 3.4.3, with addition to get $\mathbf{X}_n \Rightarrow \mathbf{X}$. (Alternatively, we could use the Skorohod representation theorem, Theorem 3.2.2.) We use the fact that addition is measurable and continuous almost surely with respect to the limit process, by virtue of the assumption about the discontinuities of $\mathbf{C}$ and $\mathbf{S}$. Specifically, for M_1 we apply Remark 12.7.1 and Theorem 12.7.3. For J_1 we apply the analog of Corollary 12.7.1; see Remark 12.6.2. Finally, we obtain the desired limit in (4.7) because

$$(\mathbf{W}_n, \mathbf{U}_n, \mathbf{L}_n) = (\phi_K(\mathbf{X}_n), \psi_U(\mathbf{X}_n), \psi_L(\mathbf{X}_n))$$

for all n. We apply the simple continuous-mapping theorem, Theorem 3.4.1, with the reflection maps, using the continuity established in Theorems 13.5.1 and 14.8.5. ■

Just as in Theorem 5.3.3, with the J_1 topology the conclusion holds in the strong SJ_1 topology as well as the product J_1 topology. As before, the conditions on the common discontinuities of $\mathbf{C}$ and $\mathbf{S}$ hold if $\mathbf{C}$ and $\mathbf{S}$ are independent processes without fixed discontinuities.

In the standard heavy-traffic applications, in addition to (4.6), we have $\lambda_n < \mu_n$, $\mu_n \to \mu$ for $0 < \mu < \infty$, $\lambda_n - \mu_n \to 0$ and $\rho_n \equiv \lambda_n/\mu_n \uparrow 1$. However, we can have non-heavy-traffic limits by having $\lambda_n n/c_n \to a > 0$ and $\mu_n n/c_n \to b > 0$, so that $c = a - b$ and $\rho_n \equiv \lambda_n/\mu_n \to a/b$, where a/b can be any positive value. Nevertheless, the heavy-traffic limit with $\rho_n \uparrow 1$ is the principal case.

We discuss heavy-traffic stochastic-process limits for the departure process and the processing time in Section 5.9. Before discussing the implications of Theorem 5.4.1, we digress to put the heavy-traffic limits in perspective with other asymptotic methods.

Remark 5.4.1. *The long tradition of asymptotics.* Given interest in the distribution of the workload $W(t)$, we perform the heavy-traffic limit, allowing $\rho_n \uparrow 1$ as $n \to \infty$ in a sequence of models index by n, to obtain simplified expressions for the ccdf $P(W(t) > x)$ and the distribution of the entire process $\{W(t) : t \geq 0\}$. We describe the resulting approximation in the Brownian case in Section 5.7 below. To put the heavy-traffic limit in perspective, we should view it in the broader context of asymptotic methods: For general mathematical models, there is a long tradition of applying asymptotic methods to obtain tractable approximations; e.g., see Bender and Orszag (1978), Bleistein and Handelsman (1986) and Olver (1974). In this tradition are the heavy-traffic approximations and asymptotic expansions obtained by Knessl and Tier (1995, 1998) using singular perturbation methods.

For stochastic processes, it is customary to perform asymptotics. We usually simplify by letting $t \to \infty$: Under regularity conditions, we obtain $W(t) \Rightarrow W(\infty)$ as $t \to \infty$ and then we focus on the limiting steady-state ccdf $P(W(\infty) > x)$. (Or, similarly, we look for a stationary distribution of the process $\{W(t) : t \geq 0\}$.) This asymptotic step is so common that it is often done without thinking. See Asmussen (1987), Baccelli and Bremaud (1994) and Borovkov (1976) for supporting theory for basic queueing processes. See Bramson (1994a,b), Baccelli and Foss (1994), Dai (1994), Meyn and Down (1994) and Borovkov (1998) for related stability results for queueing networks and more general processes.

Given a steady-state ccdf $P(W(\infty) > x)$, we may go further and let $x \to \infty$ to find the steady-state tail-probability asymptotics. As noted in Section 2.4.1, a common case for a queue with unlimited waiting space is the exponential tail:

$$P(W(\infty) > x) \sim \alpha e^{-\eta x} \quad \text{as} \quad x \to \infty ,$$

which yields the simple exponential approximation

$$P(W(\infty) > x) \approx \alpha e^{-\eta x}$$

for all x not too small; e.g., see Abate, Choudhury and Whitt (1994b, 1995).

With exponential tail-probability asymptotics, the key quantity is the asymptotic decay rate η. Since α is much less important than η, we may ignore α (i.e., let $\alpha = 1$), which corresponds to exploiting weaker large-deviation asymptotics of the form

$$\log P(W(\infty) > x) \sim -\eta x \quad \text{as} \quad x \to \infty \ ;$$

e.g., see Glynn and Whitt (1994) and Shwartz and Weiss (1995).

The large deviations limit is associated with the concept of effective bandwidths used for admission control in communication networks; see Berger and Whitt (1998a,b), Chang and Thomas (1995), Choudhury, Lucantoni and Whitt (1996), de Veciana, Kesidis and Walrand (1995), Kelly (1996) and Whitt (1993b). The idea is to assign a deterministic quantity, called the effective bandwidth, to represent how much capacity a source will require. New sources are then admitted if the sum of the effective bandwidths does not exceed the available bandwidth.

We will also consider tail-probability asymptotics applied to the steady-state distribution of the heavy-traffic limit process. We could instead consider heavy-traffic limits after establishing tail-probability asymptotics. It is significant that the two iterated limits often agree: Often the heavy-traffic asymptotics for η as $\rho \uparrow 1$ matches the asymptotics as first $t \to \infty$ and then $x \to \infty$ in the heavy-traffic limit process; see Abate and Whitt (1994b) and Choudhury and Whitt (1994). More generally, Majewski (2000) has shown that large-deviation and heavy traffic limits for queues can be interchanged. The large-deviation and heavy-traffic views are directly linked by moderate-deviations limits, which involve a different scaling, including heavy traffic ($\rho_n \uparrow 1$); see Puhalskii (1999) and Wischik (2001b).

However, as noted in Section 2.4.1, other asymptotic forms are possible for queueing processes. We often have

$$P(W(\infty) > x) \sim \alpha x^{-\beta} e^{-\eta x} \quad \text{as} \quad x \to \infty \ , \tag{4.12}$$

for non-zero β; e.g., see Abate and Whitt (1997b), Choudhury and Whitt (1996) and Duffield (1997). Moreover, even other asymptotic forms are possible; e.g., see Flatto (1997).

With heavy-tailed distributions, we usually have a power tail, i.e., (4.12) holds with $\eta = 0$:

$$P(W(\infty) > x) \sim \alpha x^{-\beta} \quad \text{as} \quad x \to \infty \ .$$

When the steady-state distribution of the workload in a queue has a power tail, the heavy-traffic theory usually is consistent; i.e., the heavy-traffic limits usually capture the relevant tail asymptotics; see Section 8.5. For more on power-tail asymptotics, see Abate, Choudhury and Whitt (1994a), Duffield and O'Connell (1995), Boxma and Dumas (1998), Sigman (1999), Jelenković (1999, 2000), Likhanov and Mazumdar (2000), Whitt (2000c) and Zwart (2000, 2001).

With the asymptotic form in (4.12), numerical transform inversion can be used to calculate the asymptotic constants η, β and α from the Laplace transform, as shown in Abate, Choudhury, Lucantoni and Whitt (1995) and Choudhury and Whitt (1996). When $\eta = 0$, we can transform the distribution into one with $\eta > 0$ to perform the computation; see Section 5 of Abate, Choudhury and Whitt (1994a) and Section 3 of Abate and Whitt (1997b). See Abate and Whitt (1996, 1999a,b,c) for ways to construct heavy-tailed distributions with tractable Laplace transforms.

And there are many other kinds of asymptotics that can be considered. For example, with queueing networks, we can let the size of the network grow; e.g., see Whitt (1984e, 1985c), Kelly (1991), Vvedenskaya et al. (1996), Mitzenmacher (1996), and Turner (1998) ■

5.5. Heavy-Traffic Scaling

A primary reason for establishing the heavy-traffic stochastic-process limit for stable queues in the previous section is to generate approximations for the workload stochastic process in a stable fluid-queue model. However, it is not exactly clear how to do this, because in applications we have one given queueing system, not a sequence of queueing systems. The general idea is to regard our given queueing system as the n^{th} queueing system in the sequence of queueing systems, but what should the value of n be?

The standard way to proceed is to choose n so that the traffic intensity ρ_n in the sequence of systems matches the actual traffic intensity in the given system. That procedure makes sense because the traffic intensity ρ is a robust first-order characterization of the system, not depending upon the stochastic fluctuations about long-term rates. As can be seen from (4.1) – (4.3) and Theorems 5.3.1 and 5.3.2, the traffic intensity appears in the fluid scaling. Thus, it is natural to think of the heavy-traffic stochastic-process limit as a way to capture the second-order variability effect beyond the traffic intensity ρ.

In controlled queueing systems, it may be necessary to solve an optimization problem to determine the relevant traffic intensity. Then the traffic intensity can not be regarded as given, but instead must be derived; see Harrison (2000, 2001a,b). After deriving the traffic intensity, we may proceed with further heavy-traffic analysis. Here we assume that the traffic intensity has been determined.

If we decide to choose n so that the traffic intensity ρ_n matches the given traffic intensity, then it is natural to index the models by the traffic intensity ρ from the outset, and then consider the limit as $\rho \uparrow 1$ (with $\uparrow$ indicating convergence upward from below). In this section we show how we can index the queueing models by the traffic intensity ρ instead of an arbitrary index n. We also discuss the applied significance of the scaling of space and time in heavy-traffic stochastic-process limits. We focus on the general fluid model considered in the last two sections, but the discussion applies to even more general models.

5.5.1. The Impact of Scaling Upon Performance

Let $W_\rho(t)$ denote the workload at time t in the infinite-capacity fluid-queue model with traffic intensity ρ. Let $c(\rho)$ and $b(\rho)$ denote the functions that scale space and time, to be identified in the next subsection. Then the scaled workload process is

$$\mathbf{W}_\rho(t) \equiv c(\rho)^{-1} W_\rho(b(\rho)t) \quad t \geq 0 . \tag{5.1}$$

The heavy-traffic stochastic-process limit can then be expressed as

$$\mathbf{W}_\rho \Rightarrow \mathbf{W} \quad \text{in} \quad (D, M_1) \quad \text{as} \quad \rho \uparrow 1 , \tag{5.2}$$

where $D \equiv D([0,\infty),\mathbb{R})$ and $\{\mathbf{W}(t) : t \geq 0\}$ is the limiting stochastic process. In the limits we consider, $c(\rho) \uparrow \infty$ and $b(\rho) \uparrow \infty$ as $\rho \uparrow 1$. Thus, the heavy-traffic stochastic-process limit provides a macroscopic view of uncertainty.

Given the heavy-traffic stochastic-process limit for the workload process in (5.2), the natural approximation is obtained by replacing the limit by approximate equality in distribution; i.e.,

$$c(\rho)^{-1} W_\rho(b(\rho)t) \approx \mathbf{W}(t), \quad t \geq 0 ,$$

or, equivalently, upon moving the scaling terms to the right side,

$$W_\rho(t) \approx c(\rho) \mathbf{W}(b(\rho)^{-1} t), \quad t \geq 0 , \tag{5.3}$$

where $\approx$ means approximately equal to in distribution (as stochastic processes).

We first discuss the applied significance of the two scaling functions $c(\rho)$ and $b(\rho)$ appearing in (5.1) and (5.3). Then, afterwards, we show how to identify these scaling functions for the fluid-queue model.

The scaling functions $c(\rho)$ and $b(\rho)$ provide important insight into queueing performance. The space-scaling factor $c(\rho)$ is relatively easy to interpret: The workload process (for times not too small) tends to be of order $c(\rho)$ as $\rho \uparrow 1$. The time-scaling factor $b(\rho)$ is somewhat more subtle: The workload process tends to make significant changes over time scales of order $b(\rho)$ as $\rho \uparrow 1$. Specifically, the change in the workload process, when adjusted for space scaling, from time $t_1 b(\rho)$ to time $t_2 b(\rho)$ is approximately characterized (for suitably high ρ) by the change in the limit process $\mathbf{W}$ from time t_1 to time t_2.

Consequently, over time intervals of length less than $b(\rho)$ the workload process tends to remain unchanged. Specifically, if we consider the change in the workload process W_ρ from time $t_1 b(\rho)$ to time $t_2(\rho)$, where $t_2(\rho) > t_1 b(\rho)$ but $t_2(\rho)/b(\rho) \to 0$ as $\rho \uparrow 1$, and if the limit process $\mathbf{W}$ is almost surely continuous at time t_1, then we conclude from the heavy-traffic limit in (5.2) that the relative change in the workload process over the time interval $[t_1 b(\rho), t_2(\rho)]$ is asymptotically negligible as ρ increases.

On the other hand, over time intervals of length greater than $b(\rho)$, the workload process W_ρ tends to approach its equilibrium steady-state distribution (assuming that both $\mathbf{W}(t)$ and $W_\rho(t)$ approach steady-state limits as $t \to \infty$). Specifically, when $t_2(\rho) > t_1 b(\rho)$ and $t_2(\rho)/b(\rho) \to \infty$ as $\rho \uparrow 1$, the workload process at time $t_2(\rho)$ tends to be in steady state, independent of its value at time $t_1 b(\rho)$. Thus, if we are considering the workload process over the time interval $[t_1 b(\rho), t_2(\rho)]$, we could use steady-state distributions to describe the distribution of $W_\rho(t_2(\rho))$, ignoring initial conditions at time $t_1 b(\rho)$. (In that step, we assume that $\mathbf{W}(t)$ approaches a steady-state distribution as $t \to \infty$, independent of initial conditions.) Thus, under regularity conditions, the time scaling in the heavy-traffic limit reveals the rate of convergence to steady state, as a function of the traffic intensity.

The use of steady-state distributions tends to be appropriate only over time intervals of length greater than $b(\rho)$. Since $b(\rho) \uparrow \infty$ as $\rho \uparrow 1$, transient (time-dependent) analysis becomes more important as ρ increases. Fortunately, the heavy-traffic stochastic-process limits provide a basis for analyzing the approximate transient behavior of the workload process as well as the approximate steady-state behavior. As indicated above, the change in the workload process (when adjusted for space scaling) between times $t_1 b(\rho)$ and $t_2 b(\rho)$ is approximately characterized by the change in the limit process $\mathbf{W}$ from time t_1 to time t_2. Fortunately, the limit processes often are sufficiently tractable that we can calculate such transient probabilities.

Remark 5.5.1. *Relaxation times.* The approximate time for a stochastic process to approach its steady-state distribution is called the *relaxation time*; e.g., see Section III.7.3 of Cohen (1982). The relaxation time can be defined in a variety of ways, but it invariably is based on the limiting behavior as $t \to \infty$ for fixed ρ. In the relatively nice light-tailed and weak-dependent case, it often can be shown, under regularity conditions, that

$$E[f(W_\rho(t))] - E[f(W_\rho(\infty))] \sim g(t,\rho)e^{-t/r(\rho)} \quad \text{as} \quad t \to \infty \ , \tag{5.4}$$

for various real-valued functions f, with the functions g and r in general depending upon f. The standard asymptotic form for the second-order term g is $g(t,\rho) \sim c(\rho)$ or $g(t,\rho) \sim c(\rho)t^{\beta(\rho)}$ as $t \to \infty$. When (5.4) holds with such a g, $r(\rho)$ is called the relaxation time. Of course, a stochastic process that starts away from steady state usually does not reach steady state in finite time. Instead, it gradually approaches steady state in a manner such as described in (5.4). More properly, we should interpret $1/r(\rho)$ as the rate of approach to steady state.

With light tails and weak dependence, we usually have

$$r(\rho)/b(\rho) \to c \quad \text{as} \quad \rho \uparrow 1 \ ,$$

where c is a positive constant; i.e., the heavy-traffic time-scaling usually reveals the asymptotic form (as $\rho \uparrow 1$) of the relaxation time.

However, with heavy tails and strong dependence, the approach to steady state is usually much slower than in (5.4); see Asmussen and Teugels (1996) and Mikosch and Nagaev (2000). In these other settings, as well as in the light-tailed weak-dependent case, the time scaling in the heavy-traffic limit usually reveals the asymptotic form (as $\rho \uparrow 1$) of the approach to steady state. Thus, the heavy-traffic time scaling can provide important insight into the rate of approach to steady state. With heavy tails and strong dependence, the heavy-traffic limits show that transient analysis becomes more important. ∎

5.5.2. Identifying Appropriate Scaling Functions

We now consider how to identify appropriate scaling functions $b(\rho)$ and $c(\rho)$ in (5.1). We can apply the general stochastic-process limit in Theorem 5.4.1 to determine appropriate scaling functions. Specifically, the scaling functions $b(\rho)$ and $c(\rho)$ depend on the input rates λ_n, the output rates μ_n

and the space-scaling factors c_n appearing in Theorem 5.4.1. The key limit is (4.6), which determines the drift η of the unreflected limit process $\mathbf{X}$.

To cover most cases of practical interest, we make *three additional assumptions* about the scaling as a function of n in (4.4): First, we assume that the space scaling is by a simple power. Specifically, we assume that

$$c_n \equiv n^H \quad \text{for} \quad 0 < H < 1 \ . \tag{5.5}$$

(See Section 4.2 for discussion about the possible scaling functions.) We need the condition on the exponent H in (5.5) in order to have $c_n \to \infty$ and $c_n/n \to 0$ as $n \to \infty$, as assumed in Theorem 5.4.1.

Second, we assume that the translation terms λ_n and μ_n in (4.4) converge to finite positive limits as $n \to \infty$. In view of condition (4.6) in Theorem 5.4.1, it suffices to assume only that

$$\mu_n \to \mu \quad \text{as} \quad n \to \infty \ , \tag{5.6}$$

where $0 < \mu < \infty$.

Third, we assume that the basic limit in (4.6) holds with $\eta < 0$. That implies that the traffic intensities ρ_n are less than 1 for all n sufficiently large. Now, if we combine (4.6), (5.5) and (5.6) (and divide by μ_n in (4.6)), we obtain the condition

$$n^{1-H}(1-\rho_n) \to \zeta \equiv -\eta/\mu > 0 \tag{5.7}$$

for $0 < \zeta < \infty$. From (5.7), we obtain the associated limit

$$n(1-\rho_n)^{1/(1-H)} \to \zeta^{1/(1-H)} \quad \text{as} \quad n \to \infty \tag{5.8}$$

or, equivalently,

$$n \sim \left(\frac{\zeta}{1-\rho_n}\right)^{\frac{1}{1-H}} \quad \text{as} \quad n \to \infty \ . \tag{5.9}$$

Thus the *canonical forms of the scaling functions* are

$$b(\rho) \equiv n \equiv \left(\frac{\zeta}{1-\rho}\right)^{\frac{1}{1-H}} \tag{5.10}$$

and

$$c(\rho) \equiv n^H \equiv \left(\frac{\zeta}{1-\rho}\right)^{\frac{H}{1-H}} \tag{5.11}$$

for $\zeta = -\eta/\mu$ as in (5.7).

To summarize, when the net-input process and potential-workload process satisfies a FCLT with time scaling by n and space scaling by n^H, the associated scaled workload processes, as functions of the traffic intensity ρ, have a heavy-traffic limit with the time-scaling function in (5.10) and space-scaling function in (5.11); i.e., as functions of ρ, the *time-scaling exponent* is $1/(1-H)$ and the *space-scaling exponent* is $H/(1-H)$.

The initial space-scaling exponent H (the Hurst parameter) depends on the burstiness; see Chapter 4. As the burstiness increases, H increases. Of course, the standard case, considered in most heavy-traffic limits for queues, is $H = 1/2$. The standard case with $H = 1/2$ occurs with Donsker's theorem and its variants with weak dependence and light tails, as discussed in Sections 4.3 and 4.4. Since $H = 1/2$ is the standard case, it is also the reference case. Values of H with $1/2 < H < 1$ indicate greater burstiness associated with heavy tails or strong positive dependence (or both). Values of H with $0 < H < 1/2$ are associated with strong negative dependence, as might occur with strong traffic shaping, e.g., scheduling.

From (5.10) and (5.11), we see that the scaling functions $b(\rho)$ and $c(\rho)$ increase rapidly as $H \uparrow 1$ for ρ near 1. Indeed, the scaling exponents increase as H increases from 0 toward 1. To make that important point clear, we display the two scaling exponents for a range of H values in Table 5.1.

H	time-scaling exponent $1/(1-H)$	space-scaling exponent $H/(1-H)$
1/101	101/100	1/100
1/11	11/10	1/10
1/5	5/4	1/4
1/3	3/2	1/2
1/2	2	1
2/3	3	2
4/5	5	4
10/11	11	10
100/101	101	100

Table 5.1: The time-scaling and space-scaling exponents as a function of the Hurst parameter H.

Since H increases as the burstiness increases, we see that increased burstiness leads to greater scaling functions $c(\rho)$ and $b(\rho)$ for any given traffic intensity ρ. The larger value of $c(\rho)$ shows that the buffer content is likely to be larger (or that one needs larger buffers to avoid overflow). The larger values of $b(\rho)$ show that the time scales for statistical regularity are longer. When there is larger burstiness, transient analysis becomes more important in contrast to steady-state analysis.

From a practical engineering perspective, the analysis of the heavy-traffic scaling functions $b(\rho)$ and $c(\rho)$ indicates that, when exceptional variability is a possibility in a queueing setting, attention should be focused on the space-scaling exponent H for the net-input process as well as the traffic intensity ρ. Second-order refinements are provided by the constant ζ appearing in (5.7), (5.10) and (5.11) and the limit process $\mathbf{W}$ appearing in (5.2) and (5.3).

5.6. Limits as the System Size Increases

In this section we see how heavy-traffic stochastic-process limits for stable fluid queues change as the system size increases. The heavy-traffic limits thus show how performance scales as the system size increases. We will see that *the performance impact depends on the way that the system size increases.* We start with a base infinite-capacity fluid queue for which there is a heavy-traffic stochastic-process limit. We assume that there is a limit for the potential-workload processes of the form $\mathbf{X}_n \Rightarrow \mathbf{X}$, where

$$\mathbf{X}_n(t) \equiv n^{-H} X_n(nt), \quad t \geq 0 , \tag{6.1}$$

for $0 < H < 1$ and

$$\mathbf{X}(t) \equiv \eta t + \mathbf{Y}(t), \quad t \geq 0 , \tag{6.2}$$

with $\{\mathbf{Y}(t) : t \geq 0\}$ being H-self-similar, i.e.,

$$\{\mathbf{Y}(ct) : t \geq 0\} \stackrel{\mathrm{d}}{=} \{c^H \mathbf{Y}(t) : t \geq 0\} \tag{6.3}$$

as in (2.5) in Section 4.2. Of course, there is a corresponding heavy-traffic stochastic-process limit for the workload process,

$$\mathbf{W}_n \Rightarrow \mathbf{W} \equiv \phi(\mathbf{X}) ,$$

where

$$\mathbf{W}_n \equiv \phi(\mathbf{X}_n) .$$

It will be convenient to focus on the potential-workload processes $\mathbf{X}_n$ instead of the workload processes $\mathbf{W}_n$. We will focus on the scale factor σ when the limit process has the representation $\mathbf{X} \equiv \eta \mathbf{e} + \sigma \mathbf{Y}$. For fixed η and $\mathbf{Y}$, the associated reflection $\{\mathbf{W}(t) : t \geq 0\}$ tends to be increasing in

σ (in a stochastic sense). For example, if $\mathbf{Y}$ is standard Brownian motion and $\eta < 0$, then the steady-state quantity $\mathbf{W}(\infty)$ has mean $\sigma^2/2|\eta|$; see (7.13) below. More generally, σ serves as a quantitative measure of the variability (for fixed $\mathbf{Y}$). The general principle is: *Increased variability in the potential workload process leads to larger workloads,* where "larger" is measured appropriately, e.g., by the mean or by a form of stochastic order.

We consider three ways to make the system larger: scaling space, scaling time and creating independent replicas. Let the *size-increase factor* be a positive integer m. We *scale space* (make it larger) by considering $m\mathbf{X}_n$; we *scale time* (make it faster) by considering $\mathbf{X}_n \circ m\mathbf{e}$; and we *create independent replicas* by considering $\mathbf{X}_{n,1} + \cdots + \mathbf{X}_{n,m}$, where $\mathbf{X}_{n,1}, \ldots, \mathbf{X}_{n,m}$ are m IID copies of the original stochastic processes $\mathbf{X}_n$.

For communication network applications, it is useful to think of constant deterministic processing, whose rate is being increased by a factor m. Scaling space then amounts to making the files or packets m times bigger to match the increased capacity. Scaling time amounts to sending the same input m times faster. Creating independent replicas means superposing (adding) m independent sources, each distributed as the original one. (We will be considering heavy-traffic limits for superposition input processes further in later chapters; see Sections 8.7.1, 9.4 and 9.8.)

In manufacturing, scaling space can also occur. Scaling space occurs in batching and unbatching; e.g., see Sections 8.5 and 9.3 of Hopp and Spearman (1996).

When we scale space, the limit process is

$$m\mathbf{X} = m\eta\mathbf{e} + m\mathbf{Y} \ . \tag{6.4}$$

When we scale time, the limit process is

$$\begin{aligned} \mathbf{X} \circ m\mathbf{e} &= m\eta\mathbf{e} + \mathbf{Y} \circ m\mathbf{e} \\ &\stackrel{\mathrm{d}}{=} m\eta\mathbf{e} + m^H\mathbf{Y} \ . \end{aligned} \tag{6.5}$$

When we create independent replicas, the limit process is

$$\sum_{i=1}^{m} \mathbf{X}_i = m\eta\mathbf{e} + \sum_{i=1}^{m} \mathbf{Y}_i \ . \tag{6.6}$$

The rate of the limit process increases by the same factor m in all three cases, but the impact on the stochastic component, characterized by the stochastic process $\mathbf{Y}$, is different for the three methods. Scaling time by m produces smaller stochastic fluctuations than scaling space by m, in the sense that the scale factors before $\mathbf{Y}$ in (6.4) and (6.5) are ordered: $m^H < m$. The advantage of time scaling over space scaling increases as H decreases (when the variability is smaller).

The impact of creating independent replicas depends on the properties of the stochastic process $\mathbf{Y}$. If $\mathbf{Y}$ is a Lévy process (has stationary and independent increments), then a concatenation of independent versions is equivalent to a longer version, i.e.,

$$\sum_{i=1}^{m} \mathbf{Y}_i \stackrel{\mathrm{d}}{=} \mathbf{Y} \circ m\mathbf{e} \ . \tag{6.7}$$

Thus, if $\mathbf{Y}$ is a Lévy process, creating independent replicas is equivalent to scaling time, which we have seen produces better performance than scaling space.

On the other hand, suppose that $\mathbf{Y}$ is fractional Brownian motion (FBM), the principal example of a non-Lévy limit process in Chapter 4. Since FBM is not a Lévy process, (6.7) does not hold. When $\mathbf{Y}$ is FBM, both $\mathbf{Y}$ and $\sum_{i=1}^{m} \mathbf{Y}_i$ are zero-mean Gaussian processes. For zero-mean Gaussian processes, it is natural to focus on the variances. With independent replicas, the variance is

$$Var \sum_{i=1}^{m} \mathbf{Y}_i(t) = m(Var\mathbf{Y}(t)), \quad t \geq 0 \ . \tag{6.8}$$

In contrast, with time scaling, because of the H-self-similarity, the variance is

$$Var\mathbf{Y}(mt) = Var(m^H \mathbf{Y}(t)) = m^{2H}(Var\mathbf{Y}(t)) \ . \tag{6.9}$$

Hence, the variance with independent replicas is less than, equal to or greater than the variance with time scaling, respectively, when $H > 1/2$, $H = 1/2$ or $H < 1/2$.

More generally, we can compare all three methods using the variance when $\mathbf{Y}(t)$ has finite variance. Using the H-self-similarity of $\mathbf{Y}$, we obtain

$$\begin{aligned} Var(m\mathbf{Y}(t)) &= m^2(Var\mathbf{Y}(t)), \\ Var\mathbf{Y}(mt) &= m^{2H}(Var\mathbf{Y}(t)), \\ Var \sum_{i=1}^{m} \mathbf{Y}_i(t) &= m(Var\mathbf{Y}(t)) \ . \end{aligned} \tag{6.10}$$

For $H < 1/2$, time scaling produces least variability; for $H > 1/2$, independent replicas produces least variability.

It is interesting to compare one large system (increased by factor m) to m separate independent systems, distributed as the original one. We say that there is *economy of scale* when the workload in the single large system tends to be smaller than the sum of the workloads in the separate systems. With finite variances, there is economy of scale when the ratio of the standard deviation to the mean is decreasing in m. From (6.10), we see that there is economy of scale with time scaling and independent replicas, but not with space scaling. For communication networks, the economy of scale associated with independent replicas is often called the *multiplexing gain*, i.e., the gain in efficiency from statistical multiplexing (combining independent sources). See Smith and Whitt (1981) for stochastic comparisons demonstrating the economy of scale in queueing systems. See Chapters 8 and 9 for more discussion.

Example 5.6.1. *Brownian motion.* Suppose that $\mathbf{X} = \eta\mathbf{e} + \sigma\mathbf{B}$, where $\eta < 0$, $\sigma > 0$ and $\mathbf{B}$ is standard Brownian motion. As noted above, the associated RBM has steady-state mean $\sigma^2/2|\eta|$. With space scaling, time scaling and creating independent replicas, the steady-state mean of the RBM's become

$$m\sigma^2/2|\eta|, \quad \sigma^2/2|\eta| \quad \text{and} \quad \sigma^2/2|\eta| \ ,$$

respectively. Thus, with space scaling, the steady-state mean is the same as the total steady-state mean in m separate systems. Otherwise, the steady-state mean is less by the factor m. ∎

In this section we have considered three different ways that the fluid queue can get larger. We have shown that the three different ways have different performance implications. It is important to realize, however, that in applications the situation may be more complicated. For example, a computer can be made larger by adding processors, but there invariably are limitations that prevent the maximum potential output rate from being proportional to the number of processors as the number of processors increases.

If the jobs are processed one at a time, then we must exploit *parallel processing*, i.e., the processors must share the processing of each job. However, usually a proportion of each job cannot be parallelized. Thus, with parallel processing, the capacity tends to increase nonlinearly with the number of processors; the marginal gain in capacity tends to be decreasing in m; e.g., see Amdahl (1967) and Chapters 5-7 and 14 of Gunther (1998). With deterministic processing, our analysis would still apply, provided that we interpret m as the actual increase in processing rate.

Even if we can accurately estimate the effective processing rate, there remain difficulties in applying the analysis in this section, because with parallel processing, it may not be appropriate to regard the processing as deterministic. It then becomes difficult to determine how the available-processing process S and its FCLT should change with m.

5.7. Brownian Approximations

In this section we apply the general heavy-traffic stochastic-process limits in Section 5.4 to establish Brownian heavy-traffic limits for fluid queues. In particular, under extra assumptions (corresponding to light tails and weak dependence), the limit for the normalized cumulative-input process will be a zero-drift Brownian motion (BM) and the limit for the normalized workload process will be a reflected Brownian motion (RBM), usually with negative drift.

The general heavy-traffic stochastic-process limits in Section 5.4 also generate non-Brownian approximations corresponding to the non-Brownian FCLT's in Chapter 4, but we do not discuss them here. We discuss approximations associated with stable Lévy motion and fractional Brownian approximations in Chapter 8.

Since Brownian motion has continuous sample paths and the reflection map maps continuous functions into continuous functions, RBM also has continuous sample paths. However, unlike Brownian motion, RBM does not have independent increments. But RBM is a Markov process. As a (well-behaved) Markov process with continuous sample paths, RBM is a diffusion process.

Harrison (1985) provides an excellent introduction to Brownian motion and "Brownian queues," showing how they can be analyzed using martingales and the Ito stochastic calculus. Other good introductions to Brownian motion and diffusion processes are Glynn (1990), Karatzas and Shreve (1988) and Chapter 15 of Karlin and Taylor (1981). Borodin and Salminen (1996) provide many Brownian formulas. Additional properties of RBM are contained in Abate and Whitt (1987a-b, 1988a-d).

5.7.1. The Brownian Limit

If $\mathbf{B}$ is a standard Brownian motion, then $\{y + \eta t + \sigma \mathbf{B}(t) : t \geq 0\}$ is a Brownian motion with *drift* η, *diffusion coefficient* (or variance coefficient) σ^2 and initial position y. We have the following elementary application of Section 5.4.

Theorem 5.7.1. (general RBM limit) *Suppose that the conditions of Theorem 5.4.1 are satisfied with $W'(0) = y$, $c_n = \sqrt{n}$ and $(\mathbf{C}, \mathbf{S})$ two-dimensional zero-drift Brownian motion with covariance matrix*

$$\Sigma = \begin{pmatrix} \sigma_C^2 & \sigma_{C,S}^2 \\ \sigma_{C,S}^2 & \sigma_S^2 \end{pmatrix} . \tag{7.1}$$

Then the conclusions of Theorems 5.4.1, 5.9.1 and 5.9.3 (b) hold with

$$(\mathbf{W}, \mathbf{U}, \mathbf{L}) \equiv (\phi_K(\mathbf{X}), \psi_U(\mathbf{X}), \psi_L(\mathbf{X}))$$

being reflected Brownian motion, i.e.,

$$\mathbf{X}(t) \stackrel{\mathrm{d}}{=} y + \eta t + \sigma_X \mathbf{B}(t) \tag{7.2}$$

for standard Brownian motion $\mathbf{B}$, drift coefficient η in (4.6) *and diffusion coefficient*

$$\sigma_X^2 = \sigma_C^2 + \sigma_S^2 - 2\sigma_{C,S}^2 . \tag{7.3}$$

Proof. Under the assumption on $(\mathbf{C}, \mathbf{S})$, $\mathbf{C} - \mathbf{S}$ is a zero-drift Brownian motion with diffusion coefficient σ_X^2 in (7.3). ∎

As indicated in Section 5.5, we can also index the queueing systems by the traffic intensity ρ and let $\rho \uparrow 1$. With $n = \zeta^2/(1-\rho)^2$ as in (5.10), the heavy-traffic limit becomes

$$\{\zeta^{-1}(1-\rho)W_\rho(t\zeta^2/(1-\rho)^2) : t \ge 0\} \Rightarrow \phi_K(\tilde{\mathbf{X}}) \quad \text{as} \quad \rho \uparrow 1 , \tag{7.4}$$

where W_ρ is the workload process in model ρ, which has output rate μ and traffic intensity ρ, and

$$\tilde{\mathbf{X}}(t) \stackrel{\mathrm{d}}{=} y - \zeta\mu t + \mathbf{B}(\sigma_X^2 t), \quad t \ge 0 , \tag{7.5}$$

with $\mathbf{B}$ being a standard Brownian motion. The capacity in model ρ is $K_\rho = \zeta K/(1-\rho)$.

We have freedom in the choice of the parameter ζ. If we let

$$\zeta = \sigma_X^2/\mu , \tag{7.6}$$

and rescale time by replacing t by t/σ_X^2, then the limit in (7.4) can be expressed as

$$\{\sigma_X^{-2}\mu(1-\rho)W_\rho(t\sigma_X^2/\mu^2(1-\rho)^2) : t \ge 0\} \Rightarrow \phi_K(\mathbf{X}) \tag{7.7}$$

where $\mathbf{X}$ is canonical Brownian motion with drift coefficient -1 and variance coeficient 1, plus initial position y, i.e.,

$$\{\mathbf{X}(t) : t \geq 0\} \stackrel{\mathrm{d}}{=} \{y - t + \mathbf{B}(t) : t \geq 0\} \ .$$

That leads to the *Brownian approximation*

$$\{W_\rho(t) : t \geq 0\} \approx \{\sigma_X^2 \mu^{-1}(1-\rho)^{-1}\phi_K(\mathbf{X})(\mu^2(1-\rho)^2 t/\sigma_X^2) : t \geq 0\} \ , \quad (7.8)$$

where $\mathbf{X}$ is again canonical Brownian motion.

Remark 5.7.1. *The impact of variability* The Brownian limit and the Brownian approximation provide insight into the way variability in the basic stochastic processes C and S affect queueing performance. In the heavy-traffic limit, the stochastic behavior of the processes C and S, beyond their rates λ and μ, affect the Brownian approximation solely via the single variance parameter σ_X^2 in (7.3), which can be identified from the CLT for $C - S$. For further discussion, see Section 9.6.1. ∎

We now show how the Brownian approximation applies to the steady-state workload.

5.7.2. The Steady-State Distribution.

The heavy-traffic limit in Theorem 5.7.1 does not directly imply that the steady-state distributions converge. Nevertheless, from (7.8), we obtain an approximation for the steady-state workload, namely,

$$W_\rho(\infty) \approx \frac{\sigma_X^2}{\mu(1-\rho)}\phi_K(\mathbf{X})(\infty) \ . \quad (7.9)$$

Conditions for the convergence of steady-state distributions in heavy traffic have been established by Szczotka (1986, 1990, 1999).

We now give the steady-state distribution of RBM with two-sided reflection; see p. 90 of Harrison (1985). We are usually interested in the case of negative drift, but we allow positive drift as well when $K < \infty$.

Theorem 5.7.2. (steady-state distribution of RBM) *Let $\{\mathbf{W}(t) : t \geq 0\}$ be one-dimensional RBM with drift coefficient η, diffusion coefficient σ^2, initial value y and two-sided reflection at 0 and K. Then*

$$\mathbf{W}(t) \Rightarrow \mathbf{W}(\infty) \quad in \quad \mathbb{R} \quad as \quad t \to \infty \ ,$$

where $\mathbf{W}(\infty)$ *has pdf*

$$f(x) \equiv \begin{cases} 1/K & \textit{if} \quad \eta = 0 \\ \\ \frac{\theta e^{\theta x}}{e^{\theta K} - 1} & \textit{if} \quad \eta \neq 0 \ , \end{cases} \tag{7.10}$$

with mean

$$E\mathbf{W}(\infty) = \begin{cases} K/2, & \textit{if} \quad \eta = 0 \\ \\ \frac{K}{1 - e^{-\theta K}} - \frac{1}{\theta} & \textit{if} \quad \eta \neq 0 \end{cases} \tag{7.11}$$

for

$$\theta \equiv 2\eta/\sigma^2 \tag{7.12}$$

Note that the steady-state distribution of RBM in (7.10) depends only on the two parameters θ in (7.12) and K. The steady-state distribution is uniform in the zero-drift case; the steady-state distribution is an exponential distribution with mean $-\theta^{-1} = \sigma^2/2|\eta|$, conditional on being in the interval $[0, K]$, when $\eta < 0$ and $\theta < 0$; $K - \mathbf{W}(\infty)$ has an exponential distribution with mean $\theta^{-1} = \sigma^2/2\eta$, conditional on being in the interval $[0, K]$, when $\eta > 0$ and $\theta > 0$. Without the upper barrier at K, a steady-state distribution exists if and only if $\eta < 0$, in which case it is the exponential distribution with mean $-\theta^{-1}$ obtained by letting $K \to \infty$ in (7.10). As K gets large, the tails of the exponential distributions rapidly become negligible so that

$$E\mathbf{W}(\infty) \approx \begin{cases} |\theta|^{-1} & \text{if} \quad \eta < 0 \\ \\ K - |\theta|^{-1} & \text{if} \quad \eta > 0 \ . \end{cases} \tag{7.13}$$

Let us now consider the approximation indicated by the limit. Since $n^{-1/2}\mathbf{W}_n(nt) \Rightarrow \mathbf{W}(t)$, we use the approximations

$$\mathbf{W}_n(t) \approx \sqrt{n}\mathbf{W}(t/n) \tag{7.14}$$

and

$$\mathbf{W}_n(\infty) \approx \sqrt{n}\mathbf{W}(\infty) \ . \tag{7.15}$$

Thus, when $K = \infty$, the Brownian approximation for $W_\rho(\infty)$ is an exponential random variable with mean

$$E[W_\rho(\infty)] \approx \frac{\sigma_X^2}{2\mu(1-\rho)} \ . \tag{7.16}$$

The RBM's $\phi_K(\tilde{\mathbf{X}})$ in (7.4) and $\phi_K(\mathbf{X})$ in (7.7) and (7.8) are the Brownian queues, which serve as the approximating models. From the approximations in (7.8) – (7.16), we see the impact upon queueing performance of the processes C and S in the heavy-traffic limit. In the heavy-traffic limit, the processes C and S affect performance through their rates $\lambda = \rho\mu$ and μ and through the variance parameter σ_X^2, which depends on the elements of the covariance matrix Σ in (7.1) as indicated in (7.3).

Note in particular that the mean of RBM in (7.16) is directly proportional to the variability of $X = C - S$ through the variability parameter σ_X^2 in (7.3). The variability parameter σ_X^2 in turn is precisely the variance constant in the CLT for the net-input process $C - S$.

In (7.9)–(7.16) we have described the approximations for the steady-state workload distribution that follow directly from the heavy-traffic limit theorem in Theorem 5.7.1. It is also possible to modify or "refine" the approximations to satisfy other criteria. For example, extra terms that appear in known exact formulas for special cases, but which are negligible in the heavy-traffic limit, may be inserted. If the goal is to develop accurate numerical approximations, then it is natural to regard heavy-traffic limits as only one of the possible theoretical reference points. For the standard multi-server GI/G/s queue, for which the heavy-traffic limit is also RBM, heuristic refinements are discussed in Whitt (1982b, 1993a) and references therein.

For the fluid queue, an important reference case for which exact formulas are available is a single-source model with independent sequences of IID on times and off times (a special case of the model studied in Chapter 8). Kella and Whitt (1992b) show that the workload process and its steady-state distribution can be related to the virtual waiting time process in the standard GI/G/1 queue (studied here in Chapter 9). Relatively simple moment formulas are thus available in the M/G/1 special case. The steady-state workload distribution can be computed in the general GI/G/1 case using numerical transform inversion, following Abate, Choudhury and Whitt (1993, 1994a, 1999). Such computations were used to illustrate the performance of bounds for general fluid queues by Choudhury and Whitt (1997).

A specific way to generate refined approximations is to interpolate between light-traffic and heavy-traffic limits; see Burman and Smith (1983, 1986), Fendick and Whitt (1989), Reiman and Simon (1988, 1989), Reiman and Weiss (1989) and Whitt (1989b). Even though numerical accuracy can be improved by refinements, the direct heavy-traffic Brownian approximations remain appealing for their simplicity.

Example 5.7.1. *The M/G/1 steady state workload.* It is instructive to compare the approximations with exact values when we can determine them. For the standard M/G/1 queue with $K = \infty$, the mean steady-state workload has the simple exact formula

$$E[W_\rho(\infty)] = \frac{\rho\sigma_X^2}{2(1-\rho)} \ , \tag{7.17}$$

which differs from (7.16) only by the factor ρ in the numerator of (7.17) and the factor μ in the denominator of (7.16). First, in the M/G/1 model the workload process has constant output rate 1, so $\mu = 1$. Hence, the only real difference between (7.16) and (7.17) is the factor ρ in the numerator of (7.17), which approaches 1 in the heavy-traffic limit.

To elaborate, in the M/G/1 queue, the cumulative input $C(t)$ equals the sum of the service times of all arrivals in the interval $[0,t]$, i.e., the cumulative input is

$$C(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \geq 0 \ ,$$

where $\{A(t) : t \geq 0\}$ is a rate-ν Poisson arrival process independent of the sequence $\{V_k : k \geq 1\}$ of IID service times, with V_1 having a general distribution with mean EV_1. Thus, the traffic intensity is $\rho \equiv \nu EV_1$. The workload process is defined in terms of the net-input process $X(t) \equiv C(t) - t$ as described in Section 5.2.

The cumulative-input process is a special case of a renewal-reward process, considered in Section 7.4. Thus, by Theorem 7.4.1, if

$$\sigma_V^2 \equiv Var V_1 < \infty \ ,$$

then the cumulative-input process obeys a FCLT $\mathbf{C}_n \Rightarrow \mathbf{C}$ for $\mathbf{C}_n$ in (3.7) with translation constant $\lambda \equiv \rho$ and space-scaling function $c_n = n^{1/2}$. Then the limit process is $\sigma_C \mathbf{B}$, where $\mathbf{B}$ is standard Brownian motion and

$$\begin{aligned} \sigma_C^2 &= \nu\sigma_V^2 + \rho EV_1 \\ &= \rho EV_1(c_V^2 + 1) \ , \end{aligned} \tag{7.18}$$

where c_V^2 is the squared coefficient of variation, defined by

$$c_V^2 \equiv \sigma_V^2/(EV_1)^2 \ . \tag{7.19}$$

Therefore,

$$\sigma_X^2 = \sigma_C^2 = \rho EV_1(c_V^2 + 1) \ . \tag{7.20}$$

With this notation, the exact formula for the mean steady-state workload in the M/G/1 queue is given in (7.17) above; e.g., see Chapter 5 of Kleinrock (1975). As indicated above, the approximation in (7.16) differs from the exact formula in (7.17) only by the factor ρ in the numerator of the exact formula, which of course disappears (becomes 1) in the heavy-traffic limit.

For the M/G/1 queue, it is known that

$$P(W_\rho(\infty) = 0) = 1 - \rho \ . \tag{7.21}$$

Thus, if we understand the approximation to be for the conditional mean $E[W_\rho(\infty)|W_\rho(\infty) > 0]$, then the approximation beomes exact. In general, however, the distribution of $W_\rho(\infty)$ is not exponential, so that the exponential distribution remains an approximation for the M/G/1 model, but the conditional distribution of $W(\infty)$ given that $W(\infty) > 0)$ is exponential in the M/M/1 special case, in which the service-time distribution is exponential. ■

5.7.3. The Overflow Process

In practice it is also of interest to describe the overflow process. In a communication network, the overflow process describes lost packets. An important design criterion is to keep the packet loss rate below a specified threshold. The *loss rate* in model n is

$$\beta_n \equiv \lim_{t\to\infty} t^{-1} U_n(t) \ . \tag{7.22}$$

The limits in Theorems 5.4.1 and 5.7.1 show that, with the heavy-traffic scaling, the loss rate should be asymptotically negligible as $n \to \infty$. Specifically, since $n^{-1/2} U_n(nt) \Rightarrow \mathbf{U}(t)$ as $n \to \infty$, where $\mathbf{U}$ is the upper-barrier regulator process of RBM, the cumulative loss in the interval $[0, n]$ is of order $\sqrt{n}$, so that the loss rate should be of order $1/\sqrt{n}$ as $n \to \infty$. (Of course, this asymptotic form depends on having the upper barriers grow as $K_n = \sqrt{n}K$ and $\rho_n \to 1$.) More precisely, we approximate the loss rate β_n by

$$\beta_n \approx \beta/\sqrt{n} \ , \tag{7.23}$$

where

$$\beta \equiv \lim_{t\to\infty} t^{-1} \mathbf{U}(t) \ . \tag{7.24}$$

Note that approximation (7.23) involves an unjustified interchange of limits, involving $n \to \infty$ and $t \to \infty$.

Berger and Whitt (1992b) make numerical comparisons (based on exact numerical algorithms) showing how the Brownian approximation in (7.23) performs for finite-capacity queues. For very small loss rates, such as 10^{-9}, it is not possible to achieve high accuracy. (Systems with the same heavy-traffic limit may have loss rates varying from 10^{-4} to 10^{-15}.) Such very small probabilities tend to be captured better by large-deviations limits. For a simple numerical comparison, see Srikant and Whitt (2001). Overall, the Brownian approximation provides important insight. That is illustrated by the sensitivity analysis in Section 9 of Berger and Whitt (1992b).

More generally, the heavy-traffic stochastic-process limits support the approximation

$$\mathbf{U}_n(t) \approx \sqrt{n}\mathbf{U}(t/n), \quad t \geq 0 , \tag{7.25}$$

where $\mathbf{U}$ is the upper-barrier regulator process of RBM. In order for the Brownian approximation for the overflow process in (7.25) to be useful, we need to obtain useful characterizations of the upper-barrier regulator process $\mathbf{U}$ associated with RBM. It suffices to describe one of the boundary regulation processes $\mathbf{U}$ and $\mathbf{L}$, because $\mathbf{L}$ has the same structure as $\mathbf{U}$ with a drift of the opposite sign. The rates of the process $\mathbf{L}$ and $\mathbf{U}$ are determined on p. 90 of Harrison (1985).

Theorem 5.7.3. (rates of boundary regulator processes) *The rates of the boundary regulator processes exist, satisfying*

$$\alpha \equiv \lim_{t\to\infty} \frac{\mathbf{L}(t)}{t} = \lim_{t\to\infty} \frac{E\mathbf{L}(t)}{t} = \begin{cases} \sigma^2/2K & \text{if} \quad \eta = 0 \\ \frac{\eta}{e^{\theta K}-1} & \text{if} \quad \eta \neq 0 \end{cases} \tag{7.26}$$

and

$$\beta \equiv \lim_{t\to\infty} \frac{\mathbf{U}(t)}{t} = \lim_{t\to\infty} \frac{E\mathbf{U}(t)}{t} = \begin{cases} \sigma^2/2K & \text{if} \quad \eta = 0 \\ \frac{\eta}{1-e^{-\theta K}} & \text{if} \quad \eta \neq 0 \ . \end{cases} \tag{7.27}$$

It is important to note that the loss rate β depends upon the variance σ^2, either directly (when $\eta = 0$) or via θ in (7.12). We can use regenerative analysis and martingales to further describe the Brownian boundary regulation processes $\mathbf{L}$ and $\mathbf{U}$; see Berger and Whitt (1992b) and Williams (1992). Let $T_{a,b}$ be the first passage time from level a to level b within $[0, K]$. Epochs at which RBM first hits 0 after first hitting K are regeneration points for the processes $\mathbf{L}$ and $\mathbf{U}$. Assuming that the RBM starts at 0, one regeneration cycle is completed at time $T_{0,K} + T_{K,0}$. Of course, $\mathbf{L}$ increases only during $[0, T_{0,K}]$, while $\mathbf{U}$ increases only during $[T_{0,K}, T_{0,K} + T_{K,0}]$. We can apply regenerative analysis and the central limit theorem for renewal processes to show that the following limits exist

$$\alpha \equiv \lim_{t\to\infty} \frac{\mathbf{L}(t)}{t} = \lim_{t\to\infty} \frac{E\mathbf{L}(t)}{t} = \frac{E\mathbf{L}(T_{0,K} + T_{K,0})}{E(T_{0,K} + T_{K,0})} \tag{7.28}$$

$$\beta \equiv \lim_{t\to\infty} \frac{\mathbf{U}(t)}{t} = \lim_{t\to\infty} \frac{E\mathbf{U}(t)}{t} = \frac{E\mathbf{U}(T_{0,K} + T_{K,0})}{E(T_{0,K} + T_{K,0})} \tag{7.29}$$

$$\sigma_L^2 \equiv \lim_{t\to\infty} \frac{Var\,\mathbf{L}(t)}{t} \quad \text{and} \quad \sigma_U^2 \equiv \lim \frac{Var\,\mathbf{U}(t)}{t} . \tag{7.30}$$

The parameters σ_L^2 and σ_U^2 in (7.30) are the *asymptotic variance parameters* of the processes $\mathbf{L}$ and $\mathbf{U}$. It is also natural to focus on the *normalized asymptotic variance parameters*

$$c_L^2 \equiv \sigma_L^2/\alpha \quad \text{and} \quad c_U^2 \equiv \sigma_U^2/\beta . \tag{7.31}$$

Theorem 5.7.4. (normalized asymptotic variance of boundary regulator processes) *The normalized asymptotic variance parameters in* (7.31) *satisfy*

$$\begin{aligned} c_U^2 &= c_L^2 = E\left[\left(\mathbf{L}(T_{0,K}) - \frac{(T_{0,K} + T_{K,0})E\mathbf{L}(T_{0,K})}{E(T_{0,K} + T_{K,0})}\right)^2 \Big/ E\mathbf{L}(T_{0,K})\right] \\ &= \begin{cases} 2K/3 & \textit{if} \quad \eta = 0 \\ \frac{2(1-e^{2\theta K} + 4\theta K e^{\theta K})}{-\theta(1-e^{\theta K})^2} & \textit{if} \quad \eta \neq 0 \end{cases} \end{aligned} \tag{7.32}$$

for $\theta \equiv 2\eta/\sigma^2$ *as in* (7.12).

In order to obtain the last line of (7.32) in Theorem 5.7.4, and for its own sake, we use an expression for the joint transform of $\mathbf{L}(T_{0,K})$ and $T_{0,K}$ from Williams (1992). Note that it suffices to let $\sigma^2 = 1$, because if $\sigma^2 > 0$ and $\mathbf{W}$ is a $(\eta/\sigma, 1)$ RBM on $[0, K/\sigma]$, then $\sigma\mathbf{W}$ is an (η, σ^2)-RBM on $[0, K]$.

Theorem 5.7.5. (joint distribution of key variables in the regenerative representation) *For* $\sigma^2 = 1$ *and all* $s_1, s_2 \geq 0$,

$$E[\exp(-s_1\mathbf{L}(T_{0,K}) - s_2 T_{0,K})]$$

$$= \begin{cases} \frac{1}{1+s_1 K} & \textit{if} \quad \eta = 0, s_2 = 0 \\ \frac{1}{\cosh(\gamma K)+s_1\gamma^{-1}\sinh(\gamma K)} & \textit{if} \quad \eta = 0, s_2 \neq 0 \\ \frac{e^{mK}}{\cos(\gamma K)+(s_1+m)\gamma^{-1}\sinh(\gamma K)} & \textit{if} \quad \eta \neq 0 \ , \end{cases} \tag{7.33}$$

where $\gamma = \sqrt{\eta^2 + 2s_2}$.

Since an explicit expression for the Laplace transform is available, we can exploit numerical transform inversion to calculate the joint probability distribution and the marginal probability distributions of $T_{0,K}$ and $\mathbf{L}(T_{0,K})$; see Abate and Whitt (1992a, 1995a), Choudhury, Lucantoni and Whitt (1994) and Abate, Choudhury and Whitt (1999).

Explicit expressions for the moments of $\mathbf{L}(T_{0,K})$ and $T_{0,K}$ can be obtained directly from Theorem 5.7.5.

Theorem 5.7.6. (associated moments of regenerative variables) *If* $\eta = 0$ *and* $\sigma^2 = 1$, *then*

$$\begin{aligned} ET_{0,K} = K^2, \ ET_{0,K}^2 \ &= \ 5K^4/3 \ , \\ E[\mathbf{L}(T_{0,K})] = K, \ E[\mathbf{L}(T_{0,K})^2] &= 2K^2 \\ E[T_{0,K}\mathbf{L}(T_{0,K})] \ &= \ 5K^3/3 \, . \end{aligned} \tag{7.34}$$

If $\eta \neq 0$ and $\sigma^2 = 1$, then

$$\begin{aligned} ET_{0,K} &= (e^{-2\eta K} - 1 + 2\eta K)/2\eta^2 \ , \\ E[T_{0,K}^2] &= (e^{-4\eta K} + e^{-2\eta K} + 6\eta K e^{-2\eta K} + 2\eta^2 K^2 - 2)/2\eta^4 \ , \\ E[\mathbf{L}(T_{0,K})] &= (1 - e^{-2\eta K})/2\eta \ , \\ E[\mathbf{L}(T_{0,K})^2] &= (1 - e^{-2\eta K})^2/2\eta^2 \ , \\ E[T_{0,K}\mathbf{L}(T_{0,K})] &= (e^{-2\eta K} - 3\eta K e^{-2\eta K} - e^{-4\eta K} + \eta K)/2\eta^3 \ . \end{aligned} \tag{7.35}$$

Fendick and Whitt (1998) show how a Brownian approximation can be used to help interpret loss measurements in a communication network.

5.7.4. One-Sided Reflection

Even nicer descriptions of RBM are possible when there is only one reflecting barrier at the origin (corresponding to an infinite buffer). Let $\mathbf{R} \equiv \{\mathbf{R}(t;\eta,\sigma^2,x) : t \geq 0\}$ denote RBM with one reflecting barrier at the origin, i.e., $\mathbf{R} = \phi(\mathbf{B})$ for $\mathbf{B} \equiv \{\mathbf{B}(t;\eta,\sigma^2,x) : t \ge 0\}$, where ϕ is the one-dimensional reflection map in (2.5) and $\mathbf{B}$ is Brownian motion. There is a relatively simple expression for the transient distribution of RBM when there is only a single barrier; see p. 49 of Harrison (1985).

Theorem 5.7.7. (transition probability of RBM with one reflecting barrier) *If* $\mathbf{R} \equiv \{\mathbf{R}(t;\eta,\sigma^2,x) : t \ge 0\}$ *is an* (η,σ^2)*-RBM then*

$$\begin{aligned} P(\mathbf{R}(t) \le y | \mathbf{R}(0) = x) \quad &= \quad 1 - \Phi\left(\frac{-y + x + \eta t}{\sigma\sqrt{t}}\right) \\ &\qquad - \exp(2\eta y/\sigma^2)\Phi\left(\frac{-y - x - \eta t}{\sigma\sqrt{t}}\right), \end{aligned}$$

where Φ *is the standard normal cdf.*

We now observe that we can express RBM with negative drift (and one reflecting barrier at the origin) in terms of *canonical RBM* with drift coefficient -1 and diffusion coefficient 1. We first state the result for Brownian motion and then for reflected Brownian motion.

Theorem 5.7.8. (scaling to canonical Brownian motion) *If* $m < 0$ *and* $\sigma^2 > 0$, *then*

$$\{a\mathbf{B}(bt; m, \sigma^2, x) : t \ge 0\} \stackrel{\mathrm{d}}{=} \{\mathbf{B}(t; -1, 1, ax) : t \ge 0\} \tag{7.36}$$

and

$$\{\mathbf{B}(t; m, \sigma^2, x) : t \ge 0\} \stackrel{\mathrm{d}}{=} \{a^{-1}\mathbf{B}(b^{-1}t; -1, 1, ax) : t \ge 0\} \tag{7.37}$$

for

$$\begin{aligned} a \quad &= \quad \frac{|m|}{\sigma^2} > 0\,, \qquad b \;=\; \frac{\sigma^2}{m^2} > 0\,, \\ m \quad &= \quad -\frac{1}{ab} < 0\,, \qquad \sigma^2 \;=\; \frac{1}{a^2 b} > 0\,. \end{aligned} \tag{7.38}$$

Theorem 5.7.9. (scaling to canonical RBM). *If* $\eta < 0$ *and* $\sigma^2 > 0$, *then*

$$\{a\mathbf{R}(bt; \eta, \sigma^2, Y) : t \ge 0\} \stackrel{\mathrm{d}}{=} \{\mathbf{R}(t; -1, 1, aY) : t \ge 0\} \tag{7.39}$$

and

$$\{\mathbf{R}(t; \eta, \sigma^2, Y) : t \ge 0\} \stackrel{\mathrm{d}}{=} \{a^{-1}\mathbf{R}(b^{-1}t; -1, 1, aY) : t \ge 0\} \tag{7.40}$$

for

$$\begin{aligned} a &\equiv \frac{|\eta|}{\sigma^2} > 0, \quad b \equiv \frac{\sigma^2}{\eta^2}, \\ \eta &= \frac{-1}{ab}, \quad \sigma^2 = \frac{1}{a^2 b}, \end{aligned} \tag{7.41}$$

as in (7.38) *of Chapter* 4.

Theorem 5.7.9 is significant because it implies that we only need to do calculations for a single RBM — canonical RBM. Expressions for the moments of canonical RBM are given Abate and Whitt (1987a,b) along with various approximations. There it is shown that the time-dependent moments can be characterized via cdf's. In particular, the time-dependent moments starting at 0, normalized by dividing by the steady-state moments are cdf's. Moreover the differences $E(\mathbf{R}(t)|\mathbf{R}(0) = x) - E[\mathbf{R}(t)|\mathbf{R}(0) = 0]$ divided by x are complementary cdf's (ccdf's), and all these cdf's have revealing structure. Here are explicit expressions for the first two moments.

Theorem 5.7.10. (moments of canonical RBM) *If* $\mathbf{R}$ *is canonical RBM, then*

$$\begin{aligned} E[\mathbf{R}(t)|\mathbf{R}(0) = x] &= 2^{-1} + \sqrt{t}\phi\left(\frac{t-x}{\sqrt{t}}\right) \\ &- (t - x + 2^{-1})\left[1 - \Phi\left(\frac{t-x}{\sqrt{t}}\right)\right] \\ &- 2^{-1}e^{2x}\left[1 - \Phi\left(\frac{t+x}{\sqrt{t}}\right)\right] \end{aligned}$$

and

$$\begin{aligned} E[\mathbf{R}(t)^2|\mathbf{R}(0) = x] &= 2^{-1} + ((x-1)\sqrt{t} - \sqrt{t^3})\phi\left(\frac{t-x}{\sqrt{t}}\right) \\ &+ ((t-x)^2 + t - 2^{-1})\left[1 - \Phi\left(\frac{t-x}{\sqrt{t}}\right)\right] \\ &+ e^{2x}(t + x - 2^{-1})\left[1 - \Phi\left(\frac{t+x}{\sqrt{t}}\right)\right], \end{aligned}$$

where Φ *and* ϕ *are the standard normal cdf and pdf.*

When thinking about RBM approximations for queues, it is sometimes useful to regard RBM as a special M/M/1 queue with $\rho = 1$. After doing appropriate scaling, the M/M/1 queue-length process approaches a nondegenerate limit as $\rho \to 1$. Thus structure of RBM can be deduced from structure for the M/M/1 queue; see Abate and Whitt (1988a-d). This is one way to characterize the covariance function of stationary RBM; see Abate and Whitt (1988c). Recall that a nonnegative-real-valued function f is completely monotone if it has derivatives of all orders that alternate in sign. Equivalently, f can be expressed as a mixture of exponential distributions; see p. 439 of Feller (1971).

Theorem 5.7.11. (covariance function of RBM) *Let $\mathbf{R}^*$ be canonical RBM initialized by giving $\mathbf{R}^*(0)$ an exponential distribution with mean $1/2$. The process $\mathbf{R}^*$ is a stationary process with completely monotone covariance function*

$$\begin{aligned} Cov(\mathbf{R}^*(0), \mathbf{R}^*(t)) &\equiv E[\mathbf{R}^*(t) - 2^{-1})(\mathbf{R}^*(0) - 2^{-1})] \\ &= 2(1 - 2t - t^2)[1 - \Phi(\sqrt{t})] + 2\sqrt{t}(1 + t)\phi(\sqrt{t}) \\ &= H_{1e}^c(t) \ = \ H_2^c(t), \quad t \geq 0 \ , \end{aligned}$$

where H_k is the k^{th}-moment cdf and H_{1e}^c is the stationary-excess ccdf associated with the first-moment cdf, i.e.,

$$H_k(t) \equiv \frac{E[\mathbf{R}(t)^k | \mathbf{R}(0) = 0]}{E\mathbf{R}(\infty)^k}, \quad t \geq 0 \ ,$$

and

$$H_{1e}^c(t) \equiv 1 - 2\int_0^t H_1^c(s) ds, \quad t \geq 0 \ .$$

Canonical RBM has asymptotic variance

$$\sigma_{\mathbf{R}}^2 \equiv \lim_{t \to \infty} t^{-1} Var\left(\int_0^t \mathbf{R}(s) ds | \mathbf{R}(0) = x \right) = 1/2 \ .$$

5.7.5. First-Passage Times

We can also establish limits for first passage times. For a stochastic process $\{Z(t) : t \geq 0\}$, let $T_{a,b}(Z)$ denote the first passage time for Z to go from a to b. (We assume that $Z(0) = a$, and consider the first passage time to b.) In general, the first passage time functional is not continuous on D or even on the subset C, but the first passage time functional is continuous almost surely with respect to BM or RBM, because BM and RBM cross any level w.p.1 in a neighborhood of any time that they first hit a level. Hence we can invoke a version of the continuous mapping theorem to conclude that limits holds for the first passage times.

Theorem 5.7.12. (limits for first passage times) *Under the assumptions of Theorem* 5.7.1,

$$\frac{T_{a\sqrt{n},b\sqrt{n}}(W_n)}{n} \Rightarrow T_{a,b}(\mathbf{W})$$

for any positive a, b *with* $a \neq b$ *and* $0 \leq a, b \leq K$, *where* $\mathbf{W}$ *is RBM and* W_n *is the unnormalized workload process in model* n.

Now let $T_{a,b}(\mathbf{R})$ be the first-passage time from a to b for one-sided canonical RBM. The first-passage time upward is the same as when there is a (higher) upper barrier (characterized in Theorems 5.7.5 and 5.7.6), but the first-passage time down is new. Let $f(t; a, b)$ be the pdf of $T_{a,b}(\mathbf{R})$ and let $\hat{f}(s; a, b)$ be its Laplace transform, i.e.,

$$\hat{f}(s; a, b) \equiv \int_0^\infty e^{-st} f(t; a, b) dt \ ,$$

where s is a complex variable with positive real part. The Laplace transforms to and from the origin have a relatively simple form; see Abate and Whitt (1988a). Again, numerical transform inversion can be applied to compute the probability distributions themselves.

Theorem 5.7.13. (RBM first-passage-time transforms and moments) *For canonical RBM (with no upper barrier), the first-passage-time Laplace transforms to and from the origin are, respectively,*

$$\hat{f}(s; x, 0) = e^{-xr_2}$$

and

$$\hat{f}(s; 0, x) = \frac{r_1 + r_2}{r_1 e^{-xr_2} + r_2 e^{xr_1}}$$

for

$$r_1(s) = 1 + \sqrt{1 + 2s} \quad \textit{and} \quad r_2(s) = \sqrt{1 + 2s} - 1 \ ,$$

so that

$$\begin{aligned} ET_{x,0} &= x, \quad Var\, T_{x,0} \ = \ x \ , \\ ET_{0,x} &= 2^{-1}[e^{2x} - 1 - 2x] \quad \textit{and} \\ Var\, T_{0,x} &= 4^{-1}[e^{4x} - 1 - 4x + 4e^{2x}(1 - 2x) - 4] \, . \end{aligned}$$

The first passage time down is closely related to the busy period of a queue, i.e., the time from when a buffer first becomes nonempty until it becomes empty again. This concept is somewhat more complicated for fluid queues than standard queues. In either case, the distribution of the busy period for small values tends to depend on the fine structure of the model, but the tail of the busy period often can be approximated robustly, and Brownian approximations can play a useful role; see Abate and Whitt (1988d, 1995b).

First-passage-time cdf's are closely related to extreme-value ccdf's because $T_{0,a}(W) \leq t$ if and only if $W^{\uparrow}(t) \equiv \sup_{0 \leq s \leq t} W(s) \geq a$. Extreme-value theory shows that there is statistical regularity associated with both first-passage times and extreme values as $t \to \infty$ and $a \to \infty$; see Resnick (1987). Heavy-traffic extreme-value approximations for queues are discussed by Berger and Whitt (1995a), Glynn and Whitt (1995) and Chang (1997). A key limit is

$$2R^{\uparrow}(t) - \log(2t) \Rightarrow Z \quad \text{as} \quad t \to \infty ,$$

where R is canonical RBM and Z has the *Gumbel cdf*, i.e.,

$$P(Z \leq x) \equiv exp(-e^{-x}), \quad -\infty < x < \infty .$$

This limit can serve as a basis for extreme-value engineering.

To summarize, in this section we have displayed Brownian limits for a fluid queue, obtained by combining the general fluid-queue limits in Theorem 5.4.1 with the multidimensional version of Donsker's theorem in Theorem 4.3.5. We have also displayed various formulas for RBM that are helpful in applications of the Brownian limit. We discuss RBM limits and approximations further in the next section and in Sections 8.4 and 9.6.

5.8. Planning Queueing Simulations

In this section, following Whitt (1989a), we see how the Brownian approximation stemming from the Brownian heavy-traffic limit in Section 5.7 can be applied to plan simulations of queueing models. In particular, we show how the Brownian approximation can be used to estimate the required simulation run lengths needed to obtain desired statistical precision, before any data have been collected. These estimates can be used to help design the simulation experiment and even to determine whether or not a contemplated experiment should be conducted.

The queueing simulations considered are single replications (one long run) of a single queue conducted to estimate steady-state characteristics, such as long-run-average steady-state workload. For such simulations to be of genuine interest, the queueing model should be relatively complicated, so that exact numerical solution is difficult. On the other hand, the queueing model should be sufficiently tractable that we can determine an appropriate Brownian approximation.

We assume that both these criteria are met. Indeed, we specify the models that we consider by stipulating that scaled versions of the stochastic process of interest, with the standard normalization, converge to RBM as $\rho \uparrow 1$. For simplicity, we focus on the workload process in a fluid queue with infinite capacity, but the approach applies to other models as well.

Of course, such a Brownian approximation directly yields an approximation for the steady-state performance, but nevertheless we may be interested in the additional simulation in order to develop a more precise understanding of the steady-state behavior. Indeed, one use of such simulations is to evaluate how various candidate approximations perform. Then we often need to perform a large number of simulations in order to see how the approximations perform over a range of possible model parameters.

In order to exploit the Brownian approximation for a single queue, we focus on simulations of a single queue. However, the simulation actually might be for a network of queues. Then the analysis of a single queue is intended to apply to any one queue in that network. If we want to estimate the steady-state performance at all queues in the network, then the required simulation run length for the network would be the maximum required for any one queue in the network. Our analysis shows that it often suffices to focus on the bottleneck (most heavily loaded) queue in the network.

At first glance, the experimental design problem may not seem very difficult. To get a rough idea about how long the runs should be, one might do one "pilot" run to estimate the required simulation run lengths. However, such a preliminary experiment requires that you set up the entire simulation before you decide whether or not to conduct the experiment. Nevertheless, if such a sampling procedure could be employed, then the experimental design problem would indeed not be especially difficult. Interest stems from the fact that one sample run can be extremely misleading.

This queueing experimental design problem is interesting and important primarily because a uniform allocation of data over all cases (parameter values) is not nearly appropriate. Experience indicates that, for given statistical precision, the required amount of data increases as the traffic intensity increases and as the arrival-and-service variability (appropriately quantified) increases. Our goal is to quantify these phenomena.

To quantify these phenomena, we apply the space and time scaling functions. Our analysis indicates that to achieve a uniform relative error over all values of the traffic intensity ρ that the run length should be approximately proportional to the time-scaling factor $(1-\rho)^{-2}$ (for sufficiently high ρ). Relative error appears to be a good practical measure of statistical precision, except possibly when very small numbers are involved. Then absolute error might be preferred. It is interesting that the required run length depends strongly on the criterion used. With the absolute error criterion, the run length should be approximately proportional to $(1-\rho)^{-4}$. With either the relative or absolute error criteria, there obviously are great differences between the required run lengths for different values of ρ, e.g., for $\rho = 0.8$, 0.9 and 0.99.

We divide the simulation run-length problem into two components. First, there is the question: What should be the required run length given that the system starts in equilibrium (steady state)? Second, there is the question: What should we do in the customary situation in which it is not possible to start in equilibrium? We propose to delete an initial portion of each simulation run before collecting data in order to allow the system to (approximately) reach steady state. By that method, we reduce the bias (the systematic error that occurs when the expected value of the estimator differs from the quantity being estimated). The second question, then, can be restated as: How long should be the initial segment of the simulation run that is deleted?

Focusing on the first question first, we work with the workload stochastic process, assuming that we have a stationary version, denoted by W_ρ^*. First, however, note that specifying the run length has no meaning until we specify the time units. To fix the time units, we assume that the output rate in the queueing system is μ. (It usually suffices to let $\mu = 1$, but we keep general μ to show how it enters in.)

For the general fluid-queue model we have the RBM approximation in (7.8). However, since we are assuming that we start in equilibrium, instead of the Brownian approximation in (7.8), we assume that we have the associated *stationary Brownian approximation*

$$\{W_\rho^*(t) : t \geq 0\} \approx \{\sigma_X^2 \mu^{-1}(1-\rho)^{-1}\mathbf{R}^*(\sigma_X^{-2}\mu^2(1-\rho)^2 t; -1, 1) : t \geq 0\}, \tag{8.1}$$

where σ_X^2 is the variability parameter, just as in (7.8), and $\mathbf{R}^*$ is a stationary version of canonical RBM, with initial exponential distribution, i.e.,

$$\{\mathbf{R}^*(t; -1, 1) : t \geq 0\} \stackrel{\mathrm{d}}{=} \{\mathbf{R}(t; -1, 1, Y) : t \geq 0\}, \tag{8.2}$$

where the initial position Y is an exponential random variable with mean

1/2 independent of the standard Brownian motion being reflected; i.e., $\mathbf{R}^* = \phi(\mathbf{B}+Y)$ where ϕ is the reflection map and $\mathbf{B}$ is a standard Brownian motion independent of the exponential random variable Y.

The obvious application is with $\{W_\rho^*(t) : t \geq 0\}$ being a stationary version of a workload process, as defined in Section 5.2. However, our analysis applies to any stationary process having the Brownian approximation in (8.1).

5.8.1. The Standard Statistical Procedure

To describe the standard statistical procedure, let $\{W(t) : t \geq 0\}$ be a stochastic process of interest and assume that is stationary with $EW(t)^2 < \infty$. (We use that notation because we are thinking of the workload process, but the statistical procedure is more general, not even depending upon the Brownian approximation.) Our object is to estimate the mean $E[W(0)]$ by the *sample mean*, i.e., by the time average

$$\bar{W}_t \equiv t^{-1} \int_0^t W(s)ds, \quad t \geq 0 \ . \tag{8.3}$$

The standard statistical procedure, assuming ample data, is based on a CLT for $\bar{W}_t$. We assume that

$$t^{1/2}(\bar{W}_t - E[W(0)]) \Rightarrow N(0, \sigma^2) \quad \text{as} \quad t \to \infty \ , \tag{8.4}$$

where σ^2 is the *asymptotic variance*, defined by

$$\sigma^2 \equiv \lim_{t \to \infty} tVar(\bar{W}_t) = 2 \int_0^\infty C(t)dt \ , \tag{8.5}$$

and $C(t)$ is the (auto) *covariance function*

$$C(t) \equiv E[W(t)W(0)] - (E[W(0)])^2, \quad t \geq 0 \ . \tag{8.6}$$

Of course, a key part of assumption (8.4) is the requirement that the asymptotic variance σ^2 be finite. The CLT in (8.4) is naturally associated with a Brownian approximation for the process $\{W(t) : t \geq 0\}$. Such CLTs for stationary processes with weak dependence were discussed in Section 4.4. Based on (8.4), we use the normal approximation

$$\bar{W}_t \approx N(E[W(0)], \sigma^2/t) \tag{8.7}$$

for the (large) t of interest, where σ^2 is the asymptotic variance in (8.5).

Based on (8.7), a $[(1-\beta)\cdot 100]\%$ *confidence interval* for the mean $E[W(0)]$ is

$$[\bar{W}_t - z_{\beta/2}(\sigma^2/t)^{1/2},\ \bar{W}_t + z_{\beta/2}(\sigma^2/t)^{1/2}]\ , \tag{8.8}$$

where

$$P(-z_{\beta/2} \le N(0,1) \le z_{\beta/2}) = 1-\beta\ . \tag{8.9}$$

The width of the confidence interval in (8.8) provides a natural measure of the *statistical precision.* There are two natural criteria to consider: *absolute width* and *relative width.* Relative width looks at the ratio of the width to the quantity to be estimated, $E[W(0)]$.

For any given β, the absolute width and relative width of the $[(1-\beta)\cdot 100]\%$ confidence intervals for the mean $E[W(0)]$ are, respectively,

$$w_a(\beta) = \frac{2\sigma z_{\beta/2}}{t^{1/2}} \quad \text{and} \quad w_r(\beta) = \frac{2\sigma z_{\beta/2}}{t^{1/2}E[W(0)]}\ . \tag{8.10}$$

For specified *absolute width* ϵ and specified *confidence level* $1-\beta$, the required simulation run length, given (8.7), is

$$t_a(\epsilon,\beta) = \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2}\ . \tag{8.11}$$

For specified *relative width* ϵ and specified *confidence level* $1-\beta$, the required length of the estimation interval, given (8.7), is

$$t_r(\epsilon,\beta) = \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2(E[W(0)])^2}\ . \tag{8.12}$$

From (8.11) and (8.12) we draw the important and well-known conclusion that both $t_a(\epsilon,\beta)$ and $t_r(\epsilon,\beta)$ are inversely proportional to ϵ^2 and directly proportional to σ^2 and $z_{\beta/2}^2$.

Standard statistical theory describes how observations can be used to estimate the unknown quantities $E[W(0)]$ and σ^2. Instead, we apply additional information about the model to obtain rough preliminary estimates for $E[W(0)]$ and σ^2 without data.

5.8.2. Invoking the Brownian Approximation

At this point we invoke the Brownian approximation in (8.1). We assume that the process of interest is W_ρ^* and that it can be approximated by scaled stationary canonical RBM as in (8.1). The steady-state mean of canonical

RBM and its asymptotic variance are both 1/2; see Theorems 5.7.10 and 5.7.11. It thus remains to consider the scaling.

To consider the effect of scaling space and time in general, let W again be a general stationary process with covariance function C and let

$$W_{y,z}(t) \equiv yW(zt), \quad t \geq 0$$

for $y, z > 0$. Then the mean $E[W_{y,z}(t)]$, covariance function $C_{y,z}(t)$ and asymptotic variance of $W_{y,z}$ are, respectively,

$$\begin{aligned} E[W_{y,z}(t)] &= yEW(zt) = yE[W(t)] \; , \\ C_{y,z}(t) = y^2 C(zt) \quad &\text{and} \quad \sigma^2_{y,z} = y^2\sigma^2/z \; . \end{aligned} \tag{8.13}$$

Thus, from (8.1) and (8.13), we obtain the important approximations

$$E[W^*_\rho(0)] \approx \frac{\sigma^2_X}{2\mu(1-\rho)} \quad \text{and} \quad \sigma^2_{W^*_\rho} \approx \frac{\sigma^6_X}{2\mu^4(1-\rho)^4} \; . \tag{8.14}$$

We have compared the approximation for the mean in (8.14) to the exact formula for the M/G/1 workload process in Example 5.7.1. Similarly, the exact formula for the asymptotic variance for the M/M/1 workload process, where $\mu = 1$, is

$$\sigma^2_{W_\rho} = \frac{2\rho(3-\rho)}{(1-\rho)^4} \; ; \tag{8.15}$$

see (23) of Whitt (1989a). Formula (8.15) reveals limitations of the approximation in (8.14) in light traffic (as $\rho \downarrow 0$), but formula (8.15) agrees with the approximation in (8.14) in the limit as $\rho \to 1$, because $\sigma^2_X = 2\rho$ for the M/M/1 queue; let $EV_1 = 1$ and $c^2_V = 1$ in (7.20). Numerical comparisons of the predictions with simulation estimates in more general models appear in Whitt (1989a). These formulas show that the approximations give good rough approximations for ρ not too small (e.g., for $\rho \geq 1/2$).

Combining (8.12) and (8.14), we see that the approximate required simulation run length for W^*_ρ given a specified *relative* width ϵ and confidence level $1 - \beta$ for the confidence interval for $E[W^*_\rho(0)]$ is

$$t_r(\epsilon, \beta) \approx \frac{8\sigma^2_X z^2_{\beta/2}}{\epsilon^2\mu^2(1-\rho)^2} \; . \tag{8.16}$$

Combining (8.11) and (8.14), we see that the approximate required simulation run length for W^*_ρ given a specified *absolute* width ϵ and confidence level $1-\beta$ for the confidence interval for $E[W_\rho(0)]$ is

$$t_a(\epsilon,\beta) \approx \frac{2\sigma_X^6 z_{\beta/2}^2}{\epsilon^2\mu^4(1-\rho)^4} \ . \tag{8.17}$$

In summary, the Brownian approximation in (8.1) dictates that, with a criterion based on the relative width of the confidence interval, the required run length should be directly proportional to both the the time-scaling term as a function of ρ alone, $(1-\rho)^{-2}$, and the heavy-traffic variability parameter σ_X^2. In contrast, with the absolute standard error criterion, the required run length should be directly proportional to $(1-\rho)^{-4}$, the *square* of the time-scaling term as a function of ρ alone, and σ_X^6, the *cube* of the heavy-traffic variability parameter σ_X^2.

The second question mentioned at the outset is: How to determine an initial transient portion of the simulation run to delete? To develop an approximate answer, we can again apply the Brownian approximation in (8.1). If the system starts empty, we can consider canonical RBM starting empty. By Theorem 5.7.10, the time-dependent mean of canonical RBM $E[\mathbf{R}(t)|\mathbf{R}(0)=0]$ is within about 1% of its steady-state mean 1/2 at $t=4$. Hence, if we were simulating canonical RBM, then we might delete an initial portion of length 4. Thus, by (8.1), a rough rule of thumb for the queueing process W_ρ (with unit processing rate) is to delete an initial segment of length $4\sigma_X^2/\mu^2(1-\rho)^2$. When we compare this to formula (8.16), we see that the proportion of the total run that should be deleted should be about $\epsilon^2/2z_{\beta/2}^2$, which is small when ϵ is small.

We can also employ the Brownian approximation to estimate the bias due to starting away from steady-state. For example, the bias due to starting empty with canonical RBM is

$$\begin{aligned} E\bar{\mathbf{R}}_t - 1/2 &= t^{-1}\int_0^t (E[\mathbf{R}(t;-1,1,0]-1/2)ds \\ &\approx t^{-1}\int_0^\infty (E[\mathbf{R}(s);-1,1,0]-1/2)ds = 1/4t \ , \end{aligned} \tag{8.18}$$

by Corollary 1.3.4 of Abate and Whitt (1987a). The approximate relative bias is thus $1/2t$. That same relative bias should apply approximately to the workload process in the queue. We can also estimate the reduced bias due to deleting an initial portion of the run, using Theorem 5.7.10 and the hyperexponential approximation

$$1/2 - E[\mathbf{R}(t;-1,1,0] \approx 0.36e^{-5.23t} + 0.138e^{-0.764t}, \quad t \geq 0 \ . \tag{8.19}$$

Our entire analysis depends on the normal approximation in (8.7), which in turn depends on the simulation run length t. Not only must t be sufficiently large so that the estimated statistical precision based on (8.7) is adequate, but t must be sufficiently large so that the normal approximation in (8.7) is itself reasonable. Consistent with intuition, experience indicates that the run length required for (8.7) to be a reasonable approximation also depends on the parameters ρ and σ_X^2, with t needing to increase as ρ and σ_X^2 increase. We can again apply the Brownian approximation to estimate the run length required. We can ask what run length is appropriate for a normal approximation to the distribution of the sample mean of canonical RBM. First, however, the time scaling alone tells us that the run length must be at least of order $\sigma_X^2/\mu^2(1-\rho)^2$. This rough analysis indicates that the requirement for (8.7) to be a reasonable approximation is approximately the same as the requirement to control the relative standard error. For further analysis supporting this conclusion, see Asmussen (1992).

5.9. Heavy-Traffic Limits for Other Processes

We now obtain heavy-traffic stochastic-process limits for other processes besides the workload process in the setting of Section 5.4. Specifically, we obtain limits for the departure process and the processing time.

5.9.1. The Departure Process

We first obtain limits for the departure process defined in (2.11), but in general we can have difficulties applying the continuous-mapping approach with addition starting from (2.11) because the limit processes $\mathbf{S}$ and $-\mathbf{L}$ can have common discontinuities of opposite sign. We can obtain positive results when we rule that out, again invoking Theorem 12.7.3.

Let the scaled departures processes be defined by

$$\mathbf{D}_n \equiv c_n^{-1}(D_n(nt) - \mu_n nt), \quad t \geq 0 . \tag{9.1}$$

Theorem 5.9.1. (limit for the departure process) *Let the conditions of Theorem* 5.4.1 *hold. If the topology on D is J_1, assume that* $\mathbf{S}$ *and* $\mathbf{L}$ *almost surely have no common discontinuities. If the topology on D is M_1, assume that* $\mathbf{S}$ *and* $\mathbf{L}$ *almost surely have no common discontinuities with jumps of* *common sign. Then, jointly with the limits in* (4.5) *and* (4.7),

$$\mathbf{D}_n \Rightarrow \mathbf{D} \equiv \mathbf{S} - \mathbf{L} \quad \mathit{in} \quad D \tag{9.2}$$

with the same topology, for $\mathbf{D}_n$ *in* (9.1), $\mathbf{S}$ *in* (4.5) *and* $\mathbf{L}$ *in* (4.9).

Proof. By (2.11),

$$\mathbf{D}_n = \mathbf{S}_n - \mathbf{L}_n \ .$$

By Theorem 5.4.1, $(\mathbf{S}_n, \mathbf{L}_n) \Rightarrow (\mathbf{S}, \mathbf{L})$ in D^2 jointly with the other limits. Just as in the proof of Theorem 5.4.1, we can apply the continuous mapping theorem, Theorem 3.4.3, with addition. Under the conditions on the discontinuities of $\mathbf{S}$ and $\mathbf{L}$, addition is measurable and almost surely continuous. Hence we obtain the desired limit in (9.2). ∎

The extra assumption in Theorem 5.9.1 is satisfied when $P(S_n(t) = \mu_n t, \quad t \geq 0) = 1$ or when $\mathbf{X}$ has no negative jumps (which implies that $\mathbf{L} \equiv \psi_L(\mathbf{X})$ has continuous paths).

As an alternative to (9.1), we can use the input rate λ_n in the translation term of the normalized departure process; i.e., let

$$\mathbf{D}'_n \equiv c_n^{-1}(D_n(nt) - \lambda_n nt), \quad t \geq 0 \ . \tag{9.3}$$

When the input rate appears in the translation term, we can directly compare the departure processes D_n to the cumulative-input processes C_n.

Corollary 5.9.1. (limit for the departure process with input centering) *Under the assumptions of Theorem* 5.9.1,

$$\mathbf{D}'_n \Rightarrow \mathbf{D}' \equiv -\eta\mathbf{e} + \mathbf{S} - \mathbf{L} \quad \mathit{in} \quad (D, M_1) \tag{9.4}$$

for $\mathbf{D}'_n$ *in* (9.3), η *in* (4.6), $\mathbf{e}(t) \equiv t$ *for* $t \geq 0$, $\mathbf{S}$ *in* (4.5) *and* $\mathbf{L}$ *in* (4.9).

Proof. Note that $\mathbf{D}'_n = \mathbf{D}_n - \eta_n\mathbf{e}$. Hence, as before, we can apply the continuous-mapping theorem, Theorem 3.4.3, with addition to the joint limit $(\mathbf{D}_n, \eta_n\mathbf{e}) \Rightarrow (\mathbf{D}, \eta\mathbf{e})$, which holds by virtue of Theorems 5.9.1 and 11.4.5. ∎

5.9.2. The Processing Time

We now establish heavy-traffic limits for the processing time $T(t)$ in (2.12). We first exploit (2.13) when $K = \infty$. Let the scaled processing-time processes be

$$\mathbf{T}_n(t) \equiv c_n^{-1} T_n(nt), \quad t \geq 0 \ . \tag{9.5}$$

Theorem 5.9.2. (limit for the processing time when $K = \infty$) *Suppose that, in addition to the conditions of Theorem* 5.4.1, $K = \infty$, $\mu_n \to \mu$ *as* $n \to \infty$, *where* $0 < \mu < \infty$,

$$\eta_{C,n} \equiv n(\lambda_n - \mu)/c_n \to \eta_C \tag{9.6}$$

and

$$\eta_{S,n} \equiv n(\mu_n - \mu)/c_n \to \eta_S \ , \tag{9.7}$$

where $-\infty < \eta_C < \infty$ *and* $-\infty < \eta_S < \infty$, *so that* $\eta = \eta_C - \eta_S$. *If the topology on* D *is* J_1, *suppose that almost surely no two of the limit processes* $\mathbf{C}$, $\mathbf{S}$ *and* $\mathbf{L}$ *have common discontinuities. If the topology on* D *is* M_1, *assume that* $\mathbf{L}$ *and* $\mathbf{C}$ *almost surely have no common discontinuities with jumps of opposite sign, and* $\mathbf{S}$ *and* $\mathbf{L}$ *almost surely have no common discontinuities with jumps of common sign. Suppose that*

$$P(\mathbf{S}(0) = 0) = 1 \ . \tag{9.8}$$

Then

$$\mathbf{T}_n \Rightarrow \mu^{-1}\mathbf{W} \quad in \quad D \tag{9.9}$$

with the same topology on D, *jointly with the limits in* (4.5) *and* (4.7), *for* $\mathbf{T}_n$ *in* (9.5) *and* $\mathbf{W}$ *in* (4.9) *with* $K = \infty$.

Proof. We can apply the continuous-mapping approach with first passage times, using the inverse map with centering in Section 13.7. Specifically, we can apply Theorem 13.7.4 with the Skorohod representation theorem, Theorem 3.2.2. From (2.13),

$$n^{-1}T_n(nt)+nt = inf\{u \ge 0 : n^{-1}S_n(nu) \ge n^{-1}(C_n(nt)+W_n'(0)+L_n(nt))\} \ .$$

By (4.5), (9.6) and (9.7),

$$(n/c_n)(\hat{\mathbf{S}}_n - \mu\mathbf{e}, \hat{\mathbf{Z}}_n - \mu\mathbf{e}) \Rightarrow (\mathbf{S} + \eta_S\mathbf{e}, \mathbf{Z} + \eta_C\mathbf{e}) \ , \tag{9.10}$$

where

$$\hat{\mathbf{S}}_n \equiv n^{-1}S_n(nt) \quad \text{and} \quad \hat{\mathbf{Z}}_n \equiv n^{-1}(C_n(nt) + W_n'(0) + L_n(nt)), \quad t \ge 0 \ .$$

We use the conditions on the discontinuities of $\mathbf{C}$ and $\mathbf{L}$ to obtain the limit

$$(n/c_n)(\hat{\mathbf{Z}}_n - \mu\mathbf{e}) \Rightarrow \mathbf{Z} + \eta_C\mathbf{e} \ ,$$

where

$$\mathbf{Z} \equiv \mathbf{C} + W'(0) + \mathbf{L} \ ,$$

by virtue of Theorem 12.7.3. Since

$$\hat{\mathbf{T}}_n(t) \equiv n^{-1}T_n(nt) = (\hat{\mathbf{S}}_n^{-1} \circ \hat{\mathbf{Z}}_n)(t) - t, \quad t \geq 0 , \tag{9.11}$$

the desired limit for $\mathbf{T}_n$ follows from Theorem 13.7.4. In particular, (9.10), (9.11) and (9.8) imply the limit

$$\begin{aligned} &(n/c_n)(\hat{\mathbf{S}}_n^{-1} \circ \hat{\mathbf{Z}}_n - \mu^{-1}\mathbf{e} \circ \mu\mathbf{e}) \\ &\quad \Rightarrow \frac{(\mathbf{Z} + \eta_C\mathbf{e}) - (\mathbf{S} + \eta_S\mathbf{e}) \circ \mu^{-1}\mathbf{e} \circ \mu\mathbf{e}}{\mu} = \frac{\mathbf{W}}{\mu} . \quad \blacksquare \end{aligned}$$

The continuity conditions in Theorem 5.9.2 are satisfied when $\mathbf{S}$ is almost surely continuous and $\mathbf{X}$ almost surely has no negative jumps (which makes $\mathbf{L}$ almost surely have continuous paths). That important case appears in the convergence to reflected stable Lévy motion in Theorem 8.5.1.

We can also obtain a FCLT for T_n when $K < \infty$ under stronger continuity conditions and pointwise convergence under weaker conditions. (It may be possible to establish analogs to part (b) below without such strong continuity conditions.)

Theorem 5.9.3. (limits for the processing time when $K \leq \infty$) *Suppose that the conditions of Theorem* 5.4.1 *hold with* $0 < K \leq \infty$ *and* $\mu_n \to \mu$, *where* $0 < \mu < \infty$.

(a) If

$$P(t \in Disc(\mathbf{S})) = P(t \in Disc(\mathbf{W})) = 0 , \tag{9.12}$$

then

$$\mathbf{T}_n(t) \Rightarrow \mu^{-1}\mathbf{W}(t) \quad in \quad \mathbb{R} . \tag{9.13}$$

(b) If

$$P(\mathbf{C} \in C) = P(\mathbf{S} \in C) = 1 , \tag{9.14}$$

then

$$\mathbf{T}_n \Rightarrow \mu^{-1}\mathbf{W} \quad in \quad (D, M_1) , \tag{9.15}$$

where $P(\mathbf{W} \in C) = 1$.

Proof. (a) By (2.12),

$$n^{-1}T_n(nt) = inf\{u \geq 0 : S_n(n(t+u)) - S_n(nt) \geq W_n(nt)\} ,$$

so that

$$\mathbf{T}_n(t) = inf\{u \geq 0 : \mu_n u + \mathbf{S}_n(t + u(c_n/n)) - \mathbf{S}_n(t) \geq \mathbf{W}_n(t)\} .$$

By the continuous-mapping approach, with condition (9.12),

$$\mathbf{T}_n(t) \Rightarrow inf\{u \geq 0 : \mu u \geq \mathbf{W}(t)\} \ ,$$

which implies the conclusion in (9.13).

(b) Under condition (9.14),

$$\sup_{0 \leq t \leq T} \{|\mathbf{S}_n(t + u(c_n/n)) - \mathbf{S}_n(t)|\} \to 0 \quad \text{as} \quad n \to \infty \quad \text{w.p.1}$$

for any T with $0 < T < \infty$; see Section 12.4. Hence the conclusion in part (a) holds uniformly over all bounded intervals. An alternative proof follows the proof of Theorem 5.9.2, including the process $\{U(t) : t \geq 0\}$ when $K < \infty$. ∎

Remark 5.9.1. *The heavy-traffic snapshot principle.* With the previous heavy-traffic theorems in this section, Theorems 5.9.2 and 5.9.3 establish a version of the heavy-traffic snapshot principle, a term coined by Reiman (1982): *In the heavy-traffic limit, the processing time is asymptotically negligible compared to the time required for the workloads to change significantly.* Since time is scaled by n, the workloads can change significantly only over time intervals of length of order n. On the other hand, since the space scaling is by c_n, where $c_n \to \infty$ but $c_n/n \to 0$ as $n \to \infty$, the workload itself tends to be only of order c_n, which is asymptotically negligible compared to n. Correspondingly, Theorems 5.9.2 and 5.9.3 show that that processing times also are of order c_n. Thus, in the heavy-traffic limit, the workload when a particle of work departs is approximately the same as the workload when that particle of work arrived.

The heavy-traffic snapshot principle also holds in queueing networks. Thus the workload seen upon each visit to a queue in the network and upon departure from the network by a particle flowing through the network is the same, in the heavy-traffic limit, as seen by that particle upon initial arrival. The heavy-traffic snapshot principle implies that network status can be communicated effectively in a heavily loaded communication network: A special packet sent from source to destination may record the buffer content at each queue on its path. Then this information may be passed back to the source by a return packet. The snapshot principle implies that the buffer contents at the queues will tend to remain near their original levels (relative to heavy-loading levels), so that the information does not become stale. (A caveat: With the fluid-limit scaling in Section 5.3, the heavy-traffic snapshot principle is not valid. In practice, we need to check if the snapshot principle applies.) For more on the impact of old information on scheduling service in queues, see Mitzenmacher (1997).

5.10. Priorities

In this book we primarily consider the standard first-come first-served (FCFS) service discipline in which input is served in order of arrival, but it can be important to consider other service disciplines to meet performance goals. We now illustrate how we can apply heavy-traffic stochastic-process limits to analyze a queue with a non-FCFS service discipline. Specifically, we now consider the fluid-queue model with priority classes. We consider the relatively tractable *preemptive-resume priority discipline*; i.e., higher-priority work immediately preempts lower-priority work and lower-priority work resumes service where it stopped when it regains access to the server. Heavy-traffic limits for the standard single-server queue with the preemptive-resume priority discipline were established by Whitt (1971a).

In general, there may be any number m of priority classes, but it suffices to consider only two because, from the perspective of any given priority class, all lower priority work can be ignored, and all higher-priority work can be lumped together. Thus, the model we consider now is the same as in Section 5.2 except that there are two priority classes. Let class 1 have priority over class 2. For $i = 1, 2$, there is a class-i cumulative-input stochastic process $\{C_i(t) : t \geq 0\}$. As before, there is a single server, a buffer with capacity K and a single service process $\{S(t) : t \geq 0\}$. (There is only a single shared buffer, not a separate buffer for each class.)

Like the polling service discipline considered in Section 2.4.2, the preemptive-resume priority service discipline is a work-conserving service policy. Thus the total workload process is the same as for the FCFS discipline considered above. We analyze the priority model to determine the performance enhancement experienced by the high-priority class and the performance degradation experienced by the low-priority class.

We first define class-i available-processing processes by letting

$$\begin{aligned} S_1(t) &\equiv S(t), \\ S_2(t) &\equiv S_1(t) - D_1(t) \ , \end{aligned} \tag{10.1}$$

where $D_1 \equiv \{D_1(t) : t \geq 0\}$ is the class-1 departure process, defined as in (2.11). We then can define the class-i potential-workload processes by

$$X_i(t) \equiv W_i(0) + C_i(t) - S_i(t) \ , \tag{10.2}$$

just as in (2.4). Then the class-i workload, overflow and departure processes are $W_i \equiv \phi_K(X_i)$, $U_i \equiv \psi_U(X_i)$ and $D_i \equiv S_i - \psi_L(X_i)$, just as in Section 5.2.

We now want to consider heavy-traffic limits for the two-priority fluid-queue model. As in Section 5.4, we consider a sequence of queues indexed by n. Suppose that the per-class input rates $\lambda_{1,n}$ and $\lambda_{2,n}$ and a maximum-potential output rate μ_n are well defined for each n, with limits as in (4.1) and (4.2). Then the class-i traffic intensity in model n is

$$\rho_{i,n} \equiv \lambda_{i,n}/\mu_n \tag{10.3}$$

and the overall traffic intensity in model n is

$$\rho_n \equiv \rho_{1,n} + \rho_{2,n} \ . \tag{10.4}$$

As a regularity condition, we suppose that $\mu_n \to \mu$ as $n \to \infty$, where $0 < \mu < \infty$.

In this context, there is some difficulty in establishing a single stochastic-process limit that generates useful approximations for both classes. It is natural to let

$$\rho_{i,n} \to \rho_i \ , \tag{10.5}$$

where $0 < \rho_i < \infty$. If we let $\rho \equiv \rho_1 + \rho_2 = 1$, then the full system is in heavy traffic, but the high-priority class is in light traffic: $\rho_{1,n} \to \rho_1 < 1$ as $n \to \infty$. That implies that the high-priority workload will be asymptotically negligible compared to the total workload in the heavy-traffic scaling. That observation is an important insight, but it does not produce useful approximations for the high-priority class.

On the other hand, if we let $\rho_1 = 1$, then the high-priority class is in heavy traffic, but $\rho \equiv \rho_1 + \rho_2 > 1$, so that the full system is unstable. Clearly, neither of these approaches is fully satisfactory. Yet another approach is to have *both* $\rho_n \to 1$ and $\rho_{1,n} \to 1$ as $n \to \infty$, but that forces $\rho_{2,n} \to 0$. Such a limit can be useful, but if the low-priority class does not contribute a small proportion of the load, then that approach will usually be unsatisfactory as well.

5.10.1. A Heirarchical Approach

What we suggest instead is a *heirarchical approach* based on considering the relevant scaling. From the scaling analysis in Section 5.5, including the time and space scaling in (5.10) and (5.11), we can see that the full system with higher traffic intensity has greater scaling than the high-priority class alone. Thus, we suggest *first* doing a heavy-traffic stochastic-process limit for the high-priority class alone, based on letting $\rho_{1,n} \uparrow 1$ and, *second*, afterwards doing a second heavy-traffic limit for both priority classes, based on fixing ρ_1 and letting $\rho_{2,n} \uparrow 1 - \rho_1$.

As a basis for these heavy-traffic limits, we assume that

$$(\mathbf{C}_{1,n}, \mathbf{C}_{2,n}, \mathbf{S}_n) \Rightarrow (\mathbf{C}_1, \mathbf{C}_2, \mathbf{S}) \tag{10.6}$$

where

$$\begin{aligned} \mathbf{C}_{1,n}(t) &\equiv n^{-H_{C,1}}(C_{1,n}(nt) - \lambda_{1,n}nt), \\ \mathbf{C}_{2,n}(t) &\equiv n^{-H_{C,2}}(C_{2,n}(nt) - \lambda_{2,n}nt), \\ \mathbf{S}_n(t) &\equiv n^{-H_S}(S_n(nt) - \mu nt) \end{aligned} \tag{10.7}$$

for $0 < H_{C,1} < 1$, $0 < H_{C,2} < 1$ and $0 < H_S < 1$. For simplicity, we let the processing rate μ be independent of n.

Note that a common case of considerable interest is the light-tailed weak-dependent case with space-scaling exponents

$$H_{C,1} = H_{C,2} = H_S = 1/2 \ , \tag{10.8}$$

but we allow other possibilities. We remark that in the light-tailed case with scaling exponents in (10.8) the heirarchical approach can be achieved directly using strong approximations; see Chen and Shen (2000). (See Section 2.2 of the Internet Supplement for a discussion of strong approximations.)

When (10.8) does not hold, then it is common for one of the three space-scaling exponents to dominate. That leads simplifications in the analysis that should be exploited. In the heavy-traffic limit, variability appears only for the processes with the largest scaling exponent.

Given a heavy-traffic stochastic-process limit as in Theorem 5.4.1 for the high-priority class alone with the space scaling factors in (10.7), we obtain the high-priority approximation

$$W_{1,\rho_1}(t) \approx \left(\frac{\zeta_1}{1-\rho_1}\right)^{\frac{H_1}{1-H_1}} \mathbf{W}_1 \left(\left(\frac{1-\rho_1}{\zeta_1}\right)^{\frac{1}{1-H_1}} t\right), \quad t \geq 0 \ , \tag{10.9}$$

as in (5.3) with the scaling functions in (5.10) and (5.11) based on the traffic intensity ρ_1 and the space-scaling exponent

$$H_1 = max\{H_{C,1}, H_S\} \ . \tag{10.10}$$

The limit process $\mathbf{W}_1$ in (10.9) is $\phi_K(\mathbf{X}_1)$ as in (4.9), where

$$\mathbf{X}_1(t) = W_1'(0) + \mathbf{C}_1(t) - \mathbf{S}(t) + \eta_1 t, \quad t \geq 0 \ ,$$

as in (4.8). If $H_{C,1} > H_S$, then $\mathbf{S}(t) = 0$ in the limit; if $H_S > H_{C,1}$, then $\mathbf{C}_1(t) = 0$ in the limit. Instead of (4.6), here we have

$$\eta_{1,n} \equiv n(\lambda_{1,n} - \mu_n)/c_n \rightarrow \eta_1 \ .$$

Next we can treat the aggregate workload of both classes using traffic intensity $\rho = \rho_1 + \rho_2$. We can think of the high-priority traffic intensity ρ_1 as fixed with $\rho_1 < 1$ and let $\rho_{2,n} \uparrow 1 - \rho_1$. By the same argument leading to (10.9), we obtain a heavy-traffic stochastic-process limit supporting the approximation

$$W_\rho(t) \approx \left(\frac{\zeta}{1-\rho}\right)^{\frac{H}{1-H}} \mathbf{W}\left(\left(\frac{1-\rho}{\zeta}\right)^{\frac{1}{1-H}} t\right), \quad t \geq 0 \ , \tag{10.11}$$

where the space-scaling exponent now is

$$H = max\{H_{C,1}, H_{C,2}, H_S\} \ . \tag{10.12}$$

The limit process $\mathbf{W}$ in (10.11) is $\phi_K(\mathbf{X})$ as in (4.9), where

$$\mathbf{X}(t) = W'(0) + \mathbf{C}_1(t) + \mathbf{C}_2(t) - \mathbf{S}(t) + \eta t, \quad t \geq 0 \ ,$$

as in (4.8). If $H_{C,i} < H$, then $\mathbf{C}_i(t) = 0$ in the limit; if $H_S < H$, then $\mathbf{S}(t) = 0$ in the limit. Instead of (4.6), here we have

$$\eta_n \equiv n(\lambda_{1,n} + \lambda_{2,n} - \mu_n)/c_n \rightarrow \eta \ .$$

Not only may the space-scaling exponent H in (10.11) differ from its counterpart H_1 in (10.9), but the parameters ρ and ζ in (10.11) routinely differ from their counterparts ρ_1 and ζ_1 in (10.9).

Of course, the low-priority workload is just the difference between the aggregate workload and the high-priority workload. If that difference is too complicated to work with, we can approximate the low-priority workload by the aggregate workload, since the high-priority workload should be relatively small, i.e.,

$$W_{2,\rho_2}(t) = W_\rho(t) - W_{1,\rho_1}(t) \approx W_\rho(t), \quad t \geq 0 \ . \tag{10.13}$$

5.10.2. Processing Times

We now consider the *per-class processing times*, i.e., the times required to complete processing of all work of that class in the system. For the high-priority class, we can apply Theorems 5.9.2 and 5.9.3 to justify (only partially when $K < \infty$) the approximation

$$T_{1,\rho_1}(t) \approx W_{1,\rho_1}(t)/\mu \ . \tag{10.14}$$

However, the low-priority processing time is more complicated because the last particle of low-priority work must wait, not only for the total aggregate workload to be processed, but also for the processing of all new high-priority work to arrive while that processing of the initial workload is going on. Nevertheless, the low-priority processing time is relatively tractable because it is the time required for the class-1 net input, starting from time t, to decrease far enough to remove the initial aggregate workload, i.e.,

$$T_2(t) \equiv inf\{u > 0 : X_1(t+u) - X_1(t) < -W(t)\} \ . \tag{10.15}$$

Note that (10.15) is essentially of the same form as (2.12). Thus, we can apply (10.15) with the reasoning in Theorem 5.9.3 to establish an analog of Theorem 5.9.3, which partly justifies the heavy-traffic approximation

$$T_{2,\rho_1,\rho_2}(t) \approx \frac{W_\rho(t)}{\mu(1-\rho_1)} \ . \tag{10.16}$$

In (10.16), $T_{2,\rho_1,\rho_2}(t)$ is the low-priority processing time as a function of the two traffic intensities and $W_\rho(t)$ is the aggregate workload at time t as a function of the total traffic intensity $\rho = \rho_1 + \rho_2$.

The heavy-traffic approximation in (10.16) should not be surprising because, as $\rho \uparrow 1$ with ρ_1 fixed, the stochastic fluctuations in X_1 should be negligible in the relatively short time required for the drift in X_1 to hit the target level; i.e., we have a separation of time scales just as in Section 2.4.2.

However, in applications, it may be important to account for the stochastic fluctuations in X_1. That is likely to be the case when ρ_1 is relatively high compared to ρ. Fortunately, the heavy-traffic limits also suggest a refined approximation. Appropriate heavy-traffic limits for X_1 alone suggest that the stochastic process $\{X_1(t) : t \geq 0\}$ can often be approximated by a Lévy process (a process with stationary and independent increments) without negative jumps. Moreover, the future net input $\{X_1(t+u) - X_1(t) : t \geq 0\}$ often can be regarded as approximately independent of $W(t)$. Under those approximating assumptions, the class-2 processing time in (10.15) becomes tractable. The Laplace transform of the conditional processing-time distribution given $W(t)$ is given on p.120 of Prabhu (1998). The conditional mean is the conditional mean in the heavy-traffic approximation in (10.16).

Remark 5.10.1. *Other service disciplines.* We conclude this section by referring to work establishing heavy-traffic limits for non-FCFS service disciplines. First, in addition to Chen and Shen (2000), Boxma, Cohen and Deng (1999) establish heavy-traffic limits for priority queues. As mentioned in Section 2.4.2, Coffman, Puhalskii and Reiman (1995, 1998), van der Mei and Levy (1997) and van der Mei (2000) establish heavy-traffic limits for polling service disciplines. Kingman (1982) showed how heavy-traffic limits can expose the behavior of a whole class of service disciplines related to random order of service. Yashkov (1993), Sengupta (1992), Grishechkin (1994), Zwart and Boxma (2000) and Boxma and Cohen (2000) establish heavy-traffic limits for the processor-sharing discipline. Fendick and Rodrigues (1991) develop a heavy-traffic approximation for the head-of-the-line generalized processor-sharing discipline. Abate and Whitt (1997a) and Limic (1999) consider the last-in first-out service discipline. Doytchinov et al. (2001) and Kruk et al. (2000) consider "real-time" queues with due dates. These alternative service disciplines are important because they significantly affect queueing performance. As we saw for the high-priority class with two priority classes, the alternative service disciplines can effectively control congestion for some customers when the input of other customers is excessive. The derivations of the heavy-traffic limits with these alternative service disciplines are fascinating because they involve quite different arguments.