Chapter 9

Single-Server Queues

9.1. Introduction

In this chapter we continue applying the continuous-mapping approach to establish heavy-traffic stochastic-process limits for stochastic processes of interest in queueing models, but now we consider the standard single-server queue, which has unlimited waiting space and the first-come first-served service discipline. That model is closely related to the infinite-capacity fluid queue considered in Chapter 5, but instead of continuous divisible fluid, customers with random service requirements arrive at random arrival times. Thus, attention here is naturally focused on integer-valued stochastic processes counting the number of arrivals, the number of departures and the number of customers in the queue.

Here is how this chapter is organized: We start in Section 9.2 by defining the basic stochastic processes in the single-server queue. Then in Section 9.3 we establish general heavy-traffic stochastic-process limits for a sequence of single-server queue models. Included among the general heavy-traffic limits for queues in Section 9.3 are limits for departure processes, so the displayed heavy-traffic limits for one queue immediately imply associated heavy-traffic limits for single-server queues in series. However, in general the limit processes for departure processes are complicated, so that the heavy-traffic limits are more useful for single queues.

In Section 9.4 and 9.5 we obtain FCLTs for arrival processes that are superpositions or splittings of other arrival processes. Along with the limits for departure processes, these results make the heavy-traffic FCLTs in Section 9.3 applicable to general acyclic open networks of single-server queues. Corresponding (more complicated) heavy-traffic limits for general single-class open networks of single-server queues (allowing feedback) are obtained in

Chapter 14 by applying the multidimensional reflection map, again with variants of the M_1 topology to treat limit processes with discontinuous sample paths.

In Section 9.6, we amplify the discussion in Section 5.7 by discussing the reflected-Brownian-motion (RBM) limit that commonly occurs in the light-tailed weak-dependent case and the associated RBM approximation that stems from it. We show how the useful functions in Chapter 13 can be applied again to yield the initial FCLTs for the arrival and service processes (required for the heavy-traffic limits) in more detailed models, e.g., where the input is from a multi-class batch-renewal process with class-dependent service times or a Markov-modulated point process.

We discuss the special case of very heavy tails in Section 9.7. When the service-time ccdf decays like $x^{-\alpha}$ as $x \to \infty$ for $0 < \alpha < 1$, the service-time mean is infinite. The queueing processes then fail to have proper steady-state distributions. The heavy-traffic stochastic-process limits are useful to describe how the queueing processes grow. Very large values then tend to be reached by jumps. The heavy-traffic limits yield useful approximations for the distributions of the time a high level is first crossed and the positions before and after that high level are crossed.

In Section 9.8 we extend the discussion in Section 8.7 by establishing heavy-traffic stochastic-process limits for single-server queues with superposition arrival processes, when the number of component arrival processes in the superposition increases in the limit. When the number of component arrival processes increases in the limit with the total rate held fixed, burstiness greater than that of a Poisson process in the component arrival processes tends to be dissipated because the superposition process approaches a Poisson process. For example, even with heavy-tailed interarrival times, the superposition process may satisfy a FCLT with a limit process having continuous sample paths. On the other hand, the limit process tends to be more complicated because it fails to have independent increments.

Finally, in Section 9.9 we discuss heuristic parametric-decomposition approximations for open queueing networks. In these approximations, arrival and service processes are each partially characterized by two parameters, one describing the rate and the other describing the variability. We show how the heavy-traffic stochastic-process limits can be used to help determine appropriate variability parameters.

9.2. The Standard Single-Server Queue

In this section we define the basic stochastic processes in the standard single-server queue, going beyond the introduction in Section 6.4.1. In this model, there is a single server and unlimited waiting space. Successive customers with random service requirements arrive at random arrival times. Upon arrival, customers wait in queue until it is their turn to receive service. Service is provided to the customers on a first-come first-served basis. After the customers enter service, they receive service without interruption and then depart.

The model can be specified by a sequence $\{(U_k, V_k) : k \geq 1\}$ of ordered pairs of nonnegative random variables. The variable U_k represents the interarrival time between customers k and $k-1$, while the variable V_k represents the service time of customer k. To fully specify the system, we also need to specify the initial conditions. For simplicity, we will assume that the first customer arrives at time $U_1 \geq 0$ to find an empty system. It is easy to extend the heavy-traffic limits to cover other initial conditions, as was done in Chapter 5.

The model here is similar to the fluid-queue model in Chapter 5, but the differences lead us to consider different processes, for which we use different notation. The main descriptive quantities of interest here are: W_k, the waiting time until beginning service for customer k; $L(t)$, the workload (in unfinished service time facing the server) at time t; $Q(t)$, the queue length (number in system, including the one in service, if any) at time t; Q_k^A, the queue length just before the k^{th} arrival; and Q_k^D, the queue length just after the k^{th} departure. The workload $L(t)$ is also the waiting time of a potential or "virtual" arrival at time t; thus the workload $L(t)$ is also called the virtual waiting time.

The waiting time of the k^{th} customer, W_k, can be expressed in terms of the waiting time of the previous customer, W_{k-1}, the service time of the previous customer, V_{k-1}, and the interarrival time between customers $k-1$ and k, U_k, by the classical *Lindley recursion*; i.e., we can make the definition

$$W_k \equiv [W_{k-1} + V_{k-1} - U_k]^+, \quad k \geq 2 \ , \tag{2.1}$$

where $[x]^+ = \max\{x, 0\}$ and $W_1 = 0$. We can apply mathematical induction to show that W_k can be expressed in terms of appropriate partial sums by a discrete-analog of the reflection map in (5.4) in Chapter 13 and in (2.5) below.

Theorem 9.2.1. *The waiting times satisfy*

$$W_k = S_k - \min\{S_j : 0 \le j \le k\}, \quad k \ge 0 \ , \tag{2.2}$$

where

$$S_k \equiv S^v_{k-1} - S^u_k, \quad S^v_k \equiv V_1 + \cdots + V_k \quad \textit{and} \quad S^u_k \equiv U_1 + \cdots + U_k, \ k \ge 1 \ ,$$

with $S_0 \equiv S^v_0 \equiv V_0 \equiv S^u_0 \equiv U_0 \equiv 0$.

Note that the indices in S^v_k in the definition of S_k are offset by 1. Nevertheless, S_k is the k^{th} partial sum

$$S_k = X_1 + \cdots + X_k, \quad k \ge 1 \ ,$$

where

$$X_i \equiv V_{i-1} - U_i, \quad i \ge 1 \ ,$$

with $V_0 \equiv 0$. Note that $W_1 = W_0 = 0$ with our definition, because $S_0 \equiv 0$ and $S_1 \le 0$ since $V_0 \equiv 0$.

We define the arrival counting process by letting

$$A(t) \equiv \max\{k \ge 0 : S^u_k \le t\}, \quad t \ge 0 \ . \tag{2.3}$$

Since S^u_k is the arrival time of customer k, $A(t)$ counts the number of arrivals in the interval $[0, t]$. We use the arrival process A to define the *cumulative-input process*. The cumulative input of work in the interval in $[0, t]$ is the sum of the service times of all arrivals in $[0, t]$, i.e., so the cumulative input can be defined as the random sum

$$C(t) \equiv S^v_{A(t)} \equiv \sum_{i=1}^{A(t)} V_i, \quad t \ge 0 \ . \tag{2.4}$$

The associated *net-input process* is

$$X(t) \equiv C(t) - t, \quad t \ge 0 \ .$$

As in the fluid queue, the workload is the one-sided reflection of X, i.e.,

$$L(t) \equiv \phi(X)(t) \equiv X(t) - \inf_{0 \le s \le t}\{X(s) \wedge 0\}, \quad t \ge 0 \ ; \tag{2.5}$$

i.e., ϕ is the reflection map in (2.5) and (2.6) in Chapter 8 and in (5.4) of Chapter 13. The workload process in the single-server queue coincides with

the workload process in the fluid queue with cumulative-input process C and deterministic processing rate 1.

Since the cumulative-input process here is a pure-jump process, the server is working if and only if the workload is positive. Thus, the *cumulative busy time* of the server is easy to express in terms of C and L, in particular,

$$B(t) \equiv C(t) - L(t), \quad t \geq 0 \ . \tag{2.6}$$

Equivalently, the cumulative idle time in $[0, t]$ is the lower-boundary regulator function associated with the reflection map, i.e.,

$$I(t) \equiv \psi_L(X)(t) \equiv - \inf_{0 \leq s \leq t} \{X(s) \wedge 0\}, \quad t \geq 0 \ , \tag{2.7}$$

and the cumulative busy time is

$$B(t) = t - I(t), \quad t \geq 0 \ . \tag{2.8}$$

Paralleling the definition of the arrival counting process A in (2.3), define a counting process associated with the service times by letting

$$N(t) \equiv \max\{k \geq 0 : S_k^v \leq t\}, \quad t \geq 0 \ . \tag{2.9}$$

Following our treatment of the waiting times and workload, we would like to think of the queue-length process $\{Q(t) : t \geq 0\}$ as the reflection of an appropriate "net-input" process. However, that is not possible in general. When the service times come from a sequence of IID exponential random variables, independent of the arrival process, we can exploit the lack of memory property of the exponential distribution to conclude that the queue-length process is distributed the same as the reflection of the process $\{A(t) - N'(t) : t \geq 0\}$, where $\{N'(t) : t \geq 0\}$ is a Poisson process counting "potential" service times. However, more generally, we do not have such a direct reflection representation, so we will have to work harder.

Let $D(t)$ count the number of departures in $[0, t]$. The *departure process* can be defined by

$$D(t) \equiv N(B(t)), \quad t \geq 0 \ . \tag{2.10}$$

We then can define the *queue-length process* by

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0 \ , \tag{2.11}$$

because we have stipulated that the first arrival finds an empty system.

Let D_k^A be the time of the k^{th} departure (the departure time of the k^{th} arrival). We can use the inverse relation for counting processes and

associated partial sums to define the *departure-time sequence* $\{D_k^A : k \geq 1\}$ in terms of the departure process $\{D(t) : t \geq 0\}$, i.e.,

$$D_k^A \equiv \inf\{t \geq 0 : D(t) \geq k\}, \quad k \geq 1 \ . \tag{2.12}$$

However, it is convenient to start with another expression for D_k^A. First, let T_k^A be the *service-start time* of customer k, with $T_0^A \equiv 0$. Since a customer must arrive and wait before starting service,

$$T_k^A \equiv S_k^u + W_k, \quad k \geq 1 \ . \tag{2.13}$$

Since service is not interrupted,

$$D_k^A = T_k^A + V_k, \quad k \geq 1 \ . \tag{2.14}$$

From Theorem 9.2.1 and (2.14), we obtain the following.

Corollary 9.2.1. (departure time representation) *The departure times satisfy*

$$D_k^A = S_k^v - \min_{0 \leq j \leq k}\{S_{j-1}^v - S_j^u\}, \qquad k \geq 1 \ . \tag{2.15}$$

We next define the continuous-time *service-start-time process* $\{T(t) : t \geq 0\}$ by letting

$$T(t) \equiv \max\{k \geq 0 : T_k^A \leq t\}, \quad t \geq 0 \ . \tag{2.16}$$

Given the continuous-time process $\{Q(t) : t \geq 0\}$, we can define the sequences $\{Q_k^A : k \geq 1\}$ and $\{Q_k^D : k \geq 1\}$ as *embedded sequences*, i.e.,

$$\begin{aligned} Q_k^A &\equiv Q(S_k^u-) \\ Q_k^D &\equiv Q(D_k^A) \ . \end{aligned} \tag{2.17}$$

Note that the definitions in (2.17) make Q_k^A the queue length before *all* arrivals at arrival epoch S_k^u, and Q_k^D is the queue length after *all* departures at departure epoch D_k^A. (Other definitions are possible, for Q_k^A if there are 0 interarrival times, and for Q_k^D if there are 0 service times).

So far, we have not introduced any probabilistic assumptions. The standard assumption is that $\{U_k : k \geq 1\}$ and $\{V_k : k \geq 1\}$ are independent sequences of IID random variables with general distributions, in which case the counting processes A and N are called renewal processes. Then, with the Kendall notation, the queueing model is called GI/GI/1, with GI denoting independence (I) with general distributions (G). The first GI refers to the interarrival times, while the second refers to the service times. The

final "1" indicates a single server. Unlimited waiting space and the FCFS service discipline are understood.

If in addition one of the distributions is exponential, deterministic, Erlang of order k (convolution of k IID exponentials) or hyperexponential of order k (mixture of k exponentials), then GI is replaced by M, D, E_k and H_k, respectively. Thus the $M/E_k/1$ model has a Poisson arrival process (associated with exponential interarrival times) with Erlang service times, while the $H_k/M/1$ model has a renewal arrival process with hyperexponential interarrival times and exponential service times. An attractive feature of the heavy-traffic limits is that they do not depend critically on the distributional assumptions or even the IID assumptions associated with the GI/GI/1 model.

9.3. Heavy-Traffic Limits

We now establish heavy-traffic stochastic-process limits for the stochastic processes in the stable single-server queue. We can obtain fluid limits (FLLN's) just as in Section 5.3, but we omit them. We go directly to the heavy-traffic limits for stable queues, as in Section 5.4.

9.3.1. The Scaled Processes

As in Section 5.4, we introduce a sequence of queueing models indexed by n. In model n, $U_{n,k}$ is the interarrival time between customers k and $k-1$, and $V_{n,k}$ is the service time of customer k. The partial sums for model n are $S_{n,k}$, $S^v_{n,k}$ and $S^u_{n,k}$, defined just as in Theorem 9.2.1 with $S_{n,0} \equiv S^u_{n,0} \equiv S^v_{n,0} \equiv 0$ for all n. The other stochastic processes are defined just as in Section 9.2, with an extra subscript n to indicate the model number.

We convert the initial model data as represented via the partial sums $S^u_{n,k}$ and $S^v_{n,k}$ into two sequences of random elements of $D \equiv D([0,\infty), \mathbb{R})$ by introducing translation and scaling, i.e., by letting

$$\begin{aligned} \mathbf{S}^u_n(t) &\equiv c_n^{-1}[S^u_{n,\lfloor nt \rfloor} - \lambda_n^{-1} nt] \ , \\ \mathbf{S}^v_n(t) &\equiv c_n^{-1}[S^v_{n,\lfloor nt \rfloor} - \mu_n^{-1} nt], \quad t \geq 0 \ , \end{aligned} \tag{3.1}$$

where λ_n, μ_n and c_n are positive constants and $\lfloor x \rfloor$ is the greatest integer less than or equal to x. We think of λ_n and μ_n in (3.1) as the arrival rate and service rate.

Since the indices of S_k^v are shifted by one, we also form the associated modification of $\mathbf{S}_n^v$ above by setting

$$\bar{\mathbf{S}}_n^v(t) \equiv c_n^{-1}[S^v_{n,\lfloor nt\rfloor -1} - \mu_n^{-1}nt], \quad t \geq 0\,, \tag{3.2}$$

where $S^v_{n,-1} \equiv 0$. (Recall that $S^v_{n,0} \equiv 0$ too.)

We then define associated random elements of D induced by the partial sums $S_{n,k}$, waiting times $W_{n,k}$, service-start times $T^A_{n,k}$ and departure times $D^A_{n,k}$ by letting

$$\begin{aligned}
\mathbf{S}_n(t) &\equiv c_n^{-1}S_{n,\lfloor nt\rfloor} = (\bar{\mathbf{S}}_n^v - \mathbf{S}_n^u)(t)\ , \\
\mathbf{W}_n(t) &\equiv c_n^{-1}W_{n,\lfloor nt\rfloor}\ , \\
\mathbf{T}_n^A(t) &\equiv c_n^{-1}[T^A_{n,\lfloor nt\rfloor} - \lambda_n^{-1}nt]\ , \\
\mathbf{D}_n^A(t) &\equiv c_n^{-1}[D^A_{n,\lfloor nt\rfloor} - \lambda_n^{-1}nt], \quad t \geq 0\ ,
\end{aligned} \tag{3.3}$$

where

$$S^v_{n,-1} \equiv S^v_{n,0} \equiv S^u_{n,0} \equiv W_{n,0} \equiv T^A_{n,0} \equiv D^A_{n,0} \equiv 0\ .$$

We next define normalized random elements of D induced by the associated continuous-time processes by letting

$$\begin{aligned}
\mathbf{A}_n(t) &\equiv c_n^{-1}[A_n(nt) - \lambda_n nt]\ , \\
\mathbf{N}_n(t) &\equiv c_n^{-1}[N_n(nt) - \mu_n nt]\ , \\
\mathbf{C}_n(t) &\equiv c_n^{-1}[C_n(nt) - \lambda_n\mu_n^{-1}nt]\ , \\
\mathbf{X}_n(t) &\equiv c_n^{-1}X_n(nt)\ , \\
\mathbf{L}_n(t) &\equiv c_n^{-1}L_n(nt)\ , \\
\mathbf{B}_n(t) &\equiv c_n^{-1}[B_n(nt) - nt]\ , \\
\mathbf{T}_n(t) &\equiv c_n^{-1}[T_n(nt) - \lambda_n nt]\ , \\
\mathbf{D}_n(t) &\equiv c_n^{-1}[D_n(nt) - \lambda_n nt]\ , \\
\mathbf{Q}_n(t) &\equiv c_n^{-1}Q_n(nt), \quad t \geq 0\ .
\end{aligned} \tag{3.4}$$

Finally, we define two sequences of random functions induced by the queue lengths at arrival epochs and departure epochs by letting

$$\begin{aligned}
\mathbf{Q}_n^A(t) &\equiv c_n^{-1}Q^A_{n,\lfloor nt\rfloor} \\
\mathbf{Q}_n^D(t) &\equiv c_n^{-1}Q^D_{n,\lfloor nt\rfloor}, \quad t \geq 0\ ,
\end{aligned} \tag{3.5}$$

where $Q^A_{n,0} \equiv Q^D_{n,0} \equiv 0$.

Notice that there are no translation terms in $\mathbf{S}_n$ and $\mathbf{W}_n$ in (3.3) or in $\mathbf{X}_n$ and $\mathbf{L}_n$ in (3.4). Thus we can apply the continuous-mapping theorem with the reflection map in (2.5) to directly obtain some initial results.

Theorem 9.3.1. (single-server-queue heavy-traffic limits directly from the reflection map) *Consider the sequence of single-server queues with the random elements in* (3.3) *and* (3.4).

(a) if

$$\mathbf{S}_n \Rightarrow \mathbf{S} \quad in \quad D$$

with the topology J_1 *or* M_1, *then*

$$\mathbf{W}_n \Rightarrow \phi(\mathbf{S}) \quad in \quad D$$

with the same topology, where ϕ *is the reflection map in* (2.5).

(b) If

$$\mathbf{X}_n \Rightarrow \mathbf{X} \quad in \quad D$$

with the topology J_1 *or* M_1, *then*

$$\mathbf{L}_n \Rightarrow \phi(\mathbf{X}) \quad in \quad D$$

with the same topology.

Proof. It follows from Theorem 9.2.1 that

$$\mathbf{W}_n = \phi(\mathbf{S}_n), \quad n \geq 1 \ ,$$

for the reflection map $\phi : D \to D$ in (2.5) and $\mathbf{S}_n$ and $\mathbf{W}_n$ in (3.3). Similarly, it follow from (2.5) that

$$\mathbf{L}_n = \phi(\mathbf{X}_n), \quad n \geq 1$$

for $\mathbf{X}_n$ and $\mathbf{L}_n$ in (3.4). Hence the stated results follow directly from the simple continuous mapping theorem, Theorem 3.4.1, because the reflection map is continuous by Theorem 13.5.1. ∎

Remark 9.3.1. *Strong and weak topologies on* D^2. Let SJ_1 and SM_1 denote the strong or standard J_1 and M_1 topologies on the product space D^k, and let WJ_1 and WM_1 denote the associated weak or product topologies on D^k. Given the limit $\mathbf{S}_n \Rightarrow \mathbf{S}$ in (D, J_1) assumed in Theorem 9.3.1 (a), we obtain the joint convergence

$$(\mathbf{S}_n, \mathbf{W}_n) \Rightarrow (\mathbf{S}, \phi(\mathbf{S})) \quad \text{in} \quad D([0,\infty), \mathbb{R}^2, SJ_1) \ . \tag{3.6}$$

However, given the same limit in (D, M_1), we do not obtain the analog of (3.6) in (D^2, SM_1). Example (14.5.1) shows that the map taking x into $(x, \phi(x))$ is *not* continuous when the range has the SM_1 topology. Hence, we use the WM_1 topology on the product space D^k. ∎

We now want to obtain limits for the random elements in (3.3), starting from convergence of the pair $(\mathbf{S}_n^u, \mathbf{S}_n^v)$ in (3.1). We start by establishing limits for the discrete-time processes.

9.3.2. Discrete-Time Processes

Before stating limits for the discrete-time processes, we establish conditions under which the two scaled service processes $\mathbf{S}_n^v$ and $\bar{\mathbf{S}}_n^v$ are asymptotically equivalent.

Theorem 9.3.2. (asymptotic equivalence of the scaled service processes) *If either* $\mathbf{S}_n^v \Rightarrow \mathbf{S}^v$ *or* $\bar{\mathbf{S}}_n^v \Rightarrow \mathbf{S}$ *in* $D([0,\infty), \mathbb{R}, \mathcal{T})$, *where* $\mathcal{T}$ *is the topology* J_1, M_1 *or* M_2, *then*

$$d_{J_1}(\mathbf{S}_n^v, \bar{\mathbf{S}}_n^v) \Rightarrow 0 \quad in \quad D([0,\infty), \mathbb{R}) \tag{3.7}$$

for d_{J_1} *in* (3.2) *in Section* 3.3 *and*

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

with the product-$\mathcal{T}$ *topology.*

Proof. Assume that $\mathbf{S}_n^v \Rightarrow \mathbf{S}^v$. (The argument is essentially the same starting with $\bar{\mathbf{S}}_n^v \Rightarrow \mathbf{S}^v$.) Use the Skorohod representation theorem to replace convergence in distribution by convergence w.p.1. By Section 12.4, the assumed convergence implies local uniform convergence at continuity points. Let t be such that $P(t \in Disc(\mathbf{S}^v)) = 0$, which holds for all but at most countably many t. By the right continuity at 0 for functions in D, the local uniform convergence holds at 0 and t. We now define homeomorphisms of $[0,t]$ needed for J_1 convergence in $D([0,t], \mathbb{R})$: Let $\nu_n : [0,t] \to [0,t]$ be defined by $\nu_n(0) = 0$, $\nu_n(t) = t$, $\nu_n(n^{-1}) = 2n^{-1}$ and $\nu_n(t - 2n^{-1}) = t - n^{-1}$ with ν_n defined by linear interpolation elsewhere. Let $\|\cdot\|_t$ be the uniform norm on $D([0,t], \mathbb{R})$. Since $\|\nu_n - \mathbf{e}\|_t = n^{-1}$ and

$$\begin{aligned} \|\bar{\mathbf{S}}_n^v \circ \nu_n - \mathbf{S}_n^v\|_t \quad &\leq \quad 2\, sup\{|\mathbf{S}_n^v(s)| : 0 \leq s \leq 2n^{-1}\} \\ &\quad +2\, sup\{|\mathbf{S}_n^v(s)| : t - n^{-1} \leq s \leq t + n^{-1}\} \\ &\quad \to 0 \quad \text{as} \quad n \to \infty\ , \end{aligned}$$

the limit in (3.7) holds in $D([0,t], \mathbb{R})$. Since such limits hold for a sequence $\{t_n\}$ with $t_n \to \infty$, we have (3.7), which implies (3.8) by Theorem 11.4.7. ∎

We apply Theorem 9.3.2 to establish heavy-traffic limits for the discrete-time processes.

Theorem 9.3.3. (heavy-traffic limits starting from arrival times and service times) *Suppose that*

$$(\mathbf{S}_n^u, \mathbf{S}_n^v) \Rightarrow (\mathbf{S}^u, \mathbf{S}^v) \quad in \quad D^2 \ , \tag{3.9}$$

where the topology is either WJ_1 *or* WM_1, $\mathbf{S}_n^u$ *and* $\mathbf{S}_n^v$ *are defined in* (3.1),

$$P(Disc(\mathbf{S}^u) \cap Disc(\mathbf{S}^v) = \phi) = 1 \ , \tag{3.10}$$

$c_n \to \infty$, $n/c_n \to \infty$, $\lambda_n^{-1} \to \lambda^{-1}$, $0 < \lambda^{-1} < \infty$, *and*

$$\eta_n \equiv n(\mu_n^{-1} - \lambda_n^{-1})/c_n \to \eta \quad as \quad n \to \infty \ . \tag{3.11}$$

(a) If the topology in (3.9) *is* WJ_1, *then*

$$(\mathbf{S}_n^u, \bar{\mathbf{S}}_n^v, \mathbf{S}_n, \mathbf{W}_n, \mathbf{T}_n^A) \Rightarrow (\mathbf{S}^u, \mathbf{S}^v, \mathbf{S}, \mathbf{W}, \mathbf{T}^A) \tag{3.12}$$

in $D([0,\infty), \mathbb{R}^5, SJ_1)$, *where*

$$\mathbf{S} \equiv \mathbf{S}^v - \mathbf{S}^u + \eta\mathbf{e}, \quad \mathbf{W} = \phi(\mathbf{S}) \quad and \quad \mathbf{T}^A = \delta(\mathbf{S}^u, \mathbf{S}^v + \eta\mathbf{e}) \ , \tag{3.13}$$

with ϕ *being the reflection map and* $\delta : D \times D \to D$ *defined by*

$$\delta(x_1, x_2) \equiv x_2 + (x_1 - x_2)^{\uparrow} \ , \tag{3.14}$$

where $x^{\uparrow}$ *is the supremum of* x *defined by*

$$x^{\uparrow}(t) \equiv \sup_{0 \le s \le t} x(s), \quad t \ge 0 \ . \tag{3.15}$$

Then the limit processes $\mathbf{S}^u$, $\mathbf{S}^v$ *and* $\mathbf{T}^A$ *have no negative jumps.*

(b) If the topology in (3.9) *above is* WM_1, *then the limit in* (3.12) *holds in* (D^5, WM_1), *with the limit processes being as in* (3.13).

Proof. We start by invoking the Skorohod representation theorem, Theorem 3.2.2, to replace the convergence in distribution in (3.9) by convergence w.p.1 for special versions. For simplicity, we do not introduce extra notation to refer to the special versions. By Theorem 9.3.2, we obtain convergence

$$(\mathbf{S}_n^u, \bar{\mathbf{S}}_n^v) \Rightarrow (\mathbf{S}^u, \mathbf{S}^v) \tag{3.16}$$

from the initial limit in (3.9), with the same topology on D^2. We then apply condition (3.10) to strengthen the convergence to be in $D([0,\infty), \mathbb{R}^2, S\mathcal{T}_1)$, drawing upon Section 12.6. Let t be any time point in $(0,\infty)$ for which

$$P(t \in Disc(\mathbf{S}^u) \cup Disc(\mathbf{S}^v)) = 0 \ .$$

There necessarily exists infinitely many such t in any bounded interval. We thus have convergence of the restrictions of $(\mathbf{S}_n^u, \bar{\mathbf{S}}_n^v)$ in $D([0,t,],\mathbb{R}^2, S\mathcal{T}_1)$, for which we again use the same notation.

(a) First, suppose that $S\mathcal{T}_1 = SJ_1$. By the definition of SJ_1 convergence, we can find increasing homeomorphisms ν_n of $[0,t]$ such that

$$\|(\mathbf{S}_n^u, \bar{\mathbf{S}}_n^v) - (\mathbf{S}^u, \mathbf{S}^v) \circ \nu_n\|_t \to 0 \quad w.p.1 \ ,$$

where $\|\cdot\|_t$ is the uniform norm on $[0,t]$. Since

$$\begin{aligned} \mathbf{S}_n &= \bar{\mathbf{S}}_n^v - \mathbf{S}_n^u + \eta_n \mathbf{e} \ , \\ \mathbf{W}_n &= \phi(\mathbf{S}_n) \ , \\ \mathbf{T}_n^A &= \delta(\mathbf{S}_n^u, \bar{\mathbf{S}}_n^v + \eta_n \mathbf{e}) \end{aligned}$$

where ϕ is the reflection map and δ is the map in (3.14), both regarded as maps from $D([0,t],\mathbb{R})$ or $D([0,t],\mathbb{R})^2$ to $D([0,t],\mathbb{R})$, which are easily seen to be Lipschitz continuous, first with respect to the uniform metric and then for d_{J_1}, it follows that

$$\|(\mathbf{S}_n^u, \bar{\mathbf{S}}_n^v, \mathbf{S}_n, \mathbf{W}_n, \mathbf{T}_n^A) - (\mathbf{S}^u, \mathbf{S}^v, \mathbf{S}, \mathbf{W}, \mathbf{T}) \circ \nu_n\|_t \to 0 \text{ w.p.1} \ ,$$

for $\mathbf{S}$, $\mathbf{W}$ and $\mathbf{T}^A$ in (3.13), so that

$$(\mathbf{S}_n^u, \bar{\mathbf{S}}_n^v, \mathbf{S}_n, \mathbf{W}_n, \mathbf{T}_n^A) \to (\mathbf{S}^u, \mathbf{S}^v, \mathbf{S}, \mathbf{W}, \mathbf{T}^A) \quad \text{in} \quad D([0,\infty), \mathbb{R}^5, SJ_1)$$

w.p.1 as claimed. Next, let J_t^+ and J_t^- be the maximum-positive-jump and maximum-negative-jump functions over $[0,t]$, i.e.,

$$J_t^+(x) \equiv \sup_{0 \le s \le t} \{x(s) - x(s-)\} \tag{3.17}$$

and

$$J_t^-(x) \equiv -\inf_{0 \le s \le t} \{x(s) - x(s-)\} \ . \tag{3.18}$$

If the topology is J_1, then the functions J_t^+ and J_t^- are continuous at all $x \in D$ for which $t \notin Disc(x)$. Since $\mathbf{S}_n^u$ and $\bar{\mathbf{S}}_n^v$ have no negative jumps, then neither do $\mathbf{S}^u$ and $\mathbf{S}^v$ if the topology limit is J_1. For any $x \in D$, $x^\uparrow \equiv \sup_{0 \le s \le t}\{x(s)\}$, $t \ge 0$, has no negative jumps. Thus, since

$$\mathbf{T}^A = \mathbf{S}^v + \eta e + (\mathbf{S}^u - \mathbf{S}^v - \eta \mathbf{e})^\uparrow \ ,$$

$\mathbf{T}^A$ also has no negative jumps when the topology is J_1.

(b) Suppose that the topology on D^2 for the convergence in (3.16) is $S\mathcal{T}_1 = SM_1$, after applying condition (3.10) to strengthen the mode of convergence from WM_1. Then we can apply the continuous maps to get the limit (3.12) with the WM_1 topology. We need the SM_1 topology on the domain in order for δ in (3.14) to be continuous. As indicated in Remark 9.3.1, unlike with the J_1 topology, we need the weaker WM_1 topology on the range product space D^k. ∎

Remark 9.3.2. *Alternative conditions.* In Theorem 9.3.3 we only use condition (3.10) to strengthen the mode of convergence in (3.9) and (3.16) to the strong topology from the weak product topology. Thus, instead of condition (3.10), we could assume that (3.9) holds with the strong topology. That could hold without (3.10) holding.

Moreover, to obtain limits for $\mathbf{S}_n$ and $\mathbf{W}_n$ with the M_1 topology, instead of (3.10), we could assume that the two limit processes $\mathbf{S}^u$ and $\mathbf{S}^v$ have no common discontinuities of common sign. Then addition is continuous by virtue of Theorem 12.7.3. However, then extra conditions would be needed to establish limits for $\mathbf{T}_n$ in (3.12) and $\mathbf{D}_n$ in Theorem 9.3.4 below. ∎

9.3.3. Continuous-Time Processes

We now establish limits for the normalized continuous-time processes in (3.4) and the embedded queue-length processes in (3.5). Now we need the M_1 topology to treat stochastic-process limits with discontinuous sample paths, because we must go from partial sums to counting processes. The limits for departure processes imply limits for queues in series and contribute to establishing limits for acyclic networks of queues.

Theorem 9.3.4. (heavy-traffic limits for continuous-time processes) *Suppose that, in addition to the conditions of Theorem* 9.3.3 (*with the topology in* (3.9) *being either* WJ_1 *or* WM_1),

$$P(\mathbf{S}^u(0) = 0) = P(\mathbf{S}^v(0) = 0) = 1 \ . \tag{3.19}$$

Then

$$\begin{aligned} &(\mathbf{A}_n, \mathbf{N}_n, \mathbf{C}_n, \mathbf{X}_n, \mathbf{L}_n, \mathbf{B}_n, \mathbf{Q}_n, \mathbf{Q}_n^A, \mathbf{Q}_n^D, \mathbf{T}_n, \mathbf{D}_n, \mathbf{D}_n^A) \\ &\Rightarrow (\mathbf{A}, \mathbf{N}, \mathbf{C}, \mathbf{X}, \mathbf{L}, \mathbf{B}, \mathbf{Q}, \mathbf{Q}^A, \mathbf{Q}^A, \mathbf{T}, \mathbf{D}, \mathbf{D}^A) \quad \textit{in } (D^{12}, WM_1) \end{aligned}$$

jointly with the limits in (3.12), *where*

$$\mathbf{A} \ \equiv \ -\lambda \mathbf{S}^u \circ \lambda \mathbf{e}, \quad \mathbf{N} \ \equiv \ -\lambda \mathbf{S}^v \circ \lambda \mathbf{e} \ ,$$

$$\begin{aligned}
\mathbf{C} &\equiv (\mathbf{S}^v - \mathbf{S}^u) \circ \lambda\mathbf{e}, \quad \mathbf{X} \equiv \mathbf{S} \circ \lambda\mathbf{e} \ , \\
\mathbf{L} &\equiv \phi(\mathbf{X}) = \mathbf{W} \circ \lambda\mathbf{e}, \quad \mathbf{B} \equiv \mathbf{X}^{\downarrow} = \mathbf{S}^{\downarrow} \circ \lambda\mathbf{e} \ , \\
\mathbf{Q} &\equiv \lambda\mathbf{L}, \quad \mathbf{Q}^A \equiv \mathbf{Q} \circ \lambda^{-1}\mathbf{e} = \lambda\mathbf{W}, \quad \mathbf{T} \equiv -\lambda\mathbf{T}^A \circ \lambda\mathbf{e} \\
\mathbf{D} &\equiv \hat{\delta}(\mathbf{A}, \mathbf{N} - \lambda^2 c\mathbf{e}), \quad \mathbf{D}^A = -\lambda^{-1}\mathbf{D} \circ \lambda^{-1}\mathbf{e}
\end{aligned} \tag{3.20}$$

where $\hat{\delta} : D \times D \to D$ *is defined by*

$$\hat{\delta}(x_1, x_2) \equiv x_2 + (x_1 - x_2)^{\downarrow} \ . \tag{3.21}$$

Proof. Again we start with the Skorohod representation theorem, Theorem 3.2.2, to replace convergence in distribution with convergence w.p.1 for the associated special versions, without introducing new notation for the special versions. By exploiting the convergence preservation of the inverse map with centering in Theorem 13.7.1 as applied to counting functions in Section 13.8, we obtain $(\mathbf{A}_n, \mathbf{N}_n) \to (\mathbf{A}, \mathbf{N})$ in (D^2, WM_1) for $(\mathbf{A}_n, \mathbf{N}_n)$ in (3.1) and $(\mathbf{A}, \mathbf{N})$ in (3.20). (We use condition (3.19) at this point.) Since $\mathbf{C}_n$ involves a random sum, we apply composition with translation as in Corollary 13.3.2 to get its limit. In particular, note that

$$\mathbf{C}_n(t) = \mathbf{S}_n^v \circ \hat{\mathbf{A}}_n(t) + \mu_n^{-1}\mathbf{A}_n(t)$$

for $\mathbf{C}_n$ and $\mathbf{A}_n$ in (3.4), where

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

Since $\mathbf{A}_n \to \mathbf{A}$, $\hat{\mathbf{A}}_n \to \lambda\mathbf{e}$. Condition (3.10) and the form of $\mathbf{A}$ in (3.20) implies that

$$P(Disc(\mathbf{A}) \cap Disc(\mathbf{S}^v \circ \lambda\mathbf{e}) = \phi) = 1 \ . \tag{3.22}$$

Thus we can apply Corollary 13.3.2 to get

$$\mathbf{C}_n \to \mathbf{S}^v \circ \lambda\mathbf{e} + \mu^{-1}(-\lambda\mathbf{S}^u \circ \lambda\mathbf{e}) = (\mathbf{S}^v - \mathbf{S}^u) \circ \lambda\mathbf{e} \ .$$

Since

$$\mathbf{X}_n(t) = \mathbf{C}_n(t) + nt(\lambda_n\mu_n^{-1} - 1)/c_n$$

and condition (3.11) holds,

$$\mathbf{X}_n \to \mathbf{C} + \lambda\eta\mathbf{e} = \mathbf{S} \circ \lambda\mathbf{e} \quad \text{in} \quad (D, M_1) \ .$$

Since $\mathbf{L}_n = \phi(\mathbf{X}_n)$, we can apply the reflection map again to treat $\mathbf{L}_n$. By (2.7) and (2.8), $\mathbf{B}_n = X_n^{\downarrow}$, where $x^{\downarrow} \equiv -(-x)^{\uparrow}$ and $x^{\uparrow}$ is the supremum map. Hence we can apply the supremum map to establish the convergence of $\mathbf{B}_n$.

The argument for the queue-length process is somewhat more complicated. The idea is to represent the random function $\mathbf{Q}_n$ as the image of the reflection map applied to an appropriate function. It turns out that we can write

$$\mathbf{Q}_n = \phi(\mathbf{A}_n - \mathbf{N}_n \circ \hat{\mathbf{B}}_n + \lambda_n \mu_n \eta_n \mathbf{e}) \ , \tag{3.23}$$

where

$$\hat{\mathbf{B}}_n(t) \equiv n^{-1} B(nt), \quad t \geq 0 \ , \tag{3.24}$$

and

$$\hat{\mathbf{B}}_n \to \mathbf{e} \quad \text{in} \quad D \quad \text{w.p.1} \tag{3.25}$$

since $\mathbf{B}_n \Rightarrow \mathbf{B}$. We can write (3.23) because

$$\begin{aligned} Q_n(t) &= A_n(t) - N_n(B_n(t)) \\ &= A_n(t) - N_n(B_n(t)) - \mu_n[t - B_n(t)] + \mu_n[t - B_n(t)] \\ &= \phi(A_n - N_n \circ B_n - \mu_n[e - B_n])(t) \ . \end{aligned} \tag{3.26}$$

The second line of (3.26) is obtained by adding and subtracting $\mu_n[t - B_n(t)]$. The third line holds because the resulting expression is equivalent to the reflection representation since $\mu_n[t - B_n(t)]$, being μ_n times the cumulative idle time in $[0, t]$, is necessarily nondecreasing and increases only when $Q_n(t) = 0$. (See Theorem 14.2.3 for more on this point.) When we introduce the scaling in the random functions, the third line of (3.26) becomes (3.23), because

$$\begin{aligned} &(\mathbf{A}_n - \mathbf{N}_n \circ \hat{\mathbf{B}}_n + \lambda_n \mu_n \eta_n \mathbf{e})(t) \\ &= c_n^{-1}(A_n(nt) - \lambda_n nt - N_n(B_n(nt)) + \mu_n B_n(nt) + (\lambda_n - \mu_n)nt) \\ &= c_n^{-1}(A_n - N_n \circ B_n - \mu_n[e - B_n])(nt), \quad t \geq 0 \ , \end{aligned} \tag{3.27}$$

and $\phi(cx \circ be) = c\phi(x) \circ be$ for all $b > 0$ and $c > 0$. We have already noted that $(\mathbf{A}_n, \mathbf{N}_n) \to (\mathbf{A}, \mathbf{N})$ in (D^2, WM_1). By (3.25) and Theorem 11.4.5, we have

$$(\mathbf{A}_n, \mathbf{N}_n, \hat{\mathbf{B}}_n) \to (\mathbf{A}, \mathbf{N}, \mathbf{e}) \quad \text{in} \quad (D^3, WM_1) \ . \tag{3.28}$$

Given (3.28), we can apply composition with Theorem 13.2.3 to get

$$(\mathbf{A}_n, \mathbf{N}_n \circ \hat{\mathbf{B}}_n) \to (\mathbf{A}, \mathbf{N}) \quad \text{in} \quad (D^2, WM_1) \ . \tag{3.29}$$

By condition (3.10) and the form of $\mathbf{A}$ and $\mathbf{N}$ in (3.20),

$$P(Disc(\mathbf{A}) \cap Disc(\mathbf{N}) = \phi) = 1 \ . \tag{3.30}$$

Hence the mode of convergence in (3.28) and (3.29) can be strengthened to SM_1. Thus, we can apply the subtraction map to get

$$\mathbf{A}_n - \mathbf{N}_n \circ \hat{\mathbf{B}}_n \to \mathbf{A} - \mathbf{N} \quad \text{in} \quad (D, M_1) \ . \tag{3.31}$$

Combining (3.23) and (3.31), we obtain

$$\mathbf{Q}_n \to \phi(\mathbf{A} - \mathbf{N} + \lambda^2 \eta \mathbf{e}) = \mathbf{Q} \quad \text{in} \quad (D, M_1) \ . \tag{3.32}$$

Next, to treat the departure processes, note that $\mathbf{D}_n = \mathbf{A}_n - \mathbf{Q}_n$. By (3.23),

$$\begin{aligned} \mathbf{D}_n &= \mathbf{N}_n \circ \hat{\mathbf{B}}_n - \lambda_n \mu_n \eta_n \mathbf{e} + (\mathbf{A}_n - \mathbf{N}_n \circ \hat{\mathbf{B}}_n + \lambda_n \mu_n \eta_n \mathbf{e})^{\downarrow} \\ &= \hat{\delta}(\mathbf{A}_n, \mathbf{N}_n \circ \hat{\mathbf{B}}_n - \lambda_n \mu_n \eta_n \mathbf{e}) \end{aligned} \tag{3.33}$$

for $\hat{\delta}$ in (3.21). Since $\hat{\delta}$ is continuous as a map from (D^2, SM_1) to (D, M_1),

$$\mathbf{D}_n \to \hat{\delta}(\mathbf{A}, \mathbf{N} - \lambda^2 \eta \mathbf{e}) \quad \text{in} \quad (D, M_1) \ .$$

We then apply the convergence-preservation property of the inverse map with centering in the context of counting functions to obtain the limit for $\mathbf{D}_n^A$ from (3.33). We can apply the composition map to treat $\mathbf{Q}_n^A$ and $\mathbf{Q}_n^D$. First, as a consequence of (3.9) and (3.11), since $n/c_n \to \infty$, we have

$$\hat{\mathbf{S}}_n^A \to \lambda^{-1}\mathbf{e} \quad \text{and} \quad \hat{\mathbf{D}}_n^A \to \lambda^{-1}\mathbf{e} \tag{3.34}$$

for

$$\hat{\mathbf{S}}_n^A(t) \equiv n^{-1} S^u_{n,\lfloor nt \rfloor} \quad \text{and} \quad \hat{\mathbf{D}}_n^A(t) \equiv n^{-1} D^A_{n,\lfloor nt \rfloor} \ . \tag{3.35}$$

Applying Theorem 11.4.5 with (3.32) and (3.34), we obtain

$$(\mathbf{Q}_n, \hat{\mathbf{S}}_n^A, \hat{\mathbf{D}}_n^D) \to (\mathbf{Q}, \lambda^{-1}\mathbf{e}, \lambda^{-1}\mathbf{e}) \quad \text{in} \quad (D^3, WM_1)$$

and, then applying Theorem 13.2.3, we obtain

$$\mathbf{Q}_n^A = \mathbf{Q}_n \circ \hat{\mathbf{S}}_n^A \to \mathbf{Q} \circ \lambda^{-1}\mathbf{e} \quad \text{and} \quad \mathbf{Q}_n^D = \mathbf{Q}_n \circ \hat{\mathbf{D}}_n^A \to \mathbf{Q} \circ \lambda^{-1}\mathbf{e}$$

in (D, M_1). Finally, limits for the normalized continuous-time service-start-time processes $\mathbf{T}_n$ follow by applying the inverse map with centering as applied to counting functions to the previous limits for $\mathbf{T}_n^A$, just as we obtained limits for $\mathbf{A}_n$ and $\mathbf{N}_n$ starting from $\mathbf{S}_n^u$ and $\mathbf{S}_n^v$. ∎

Remark 9.3.3. *Impossibility of improving from M_1 to J_1.* When the limit processes have discontinuous sample paths, the M_1 mode of convergence in Theorem 9.3.4 cannot be improved to J_1. First, it is known that $\mathbf{S}_n^u \Rightarrow \mathbf{S}^u$

and $\mathbf{A}_n \Rightarrow -\lambda \mathbf{S}^u \circ \lambda e$ both hold in (D, J_1) if and only if $P(\mathbf{S}^u \in C) = 1$; see Lemma 13.7.1. In the special case of identical deterministic service times, the processes $\mathbf{C}_n$, $\mathbf{X}_n$, $\mathbf{L}_n$ and $\mathbf{Q}_n$ are simple functions of $\mathbf{A}_n$, so their convergence also cannot be strengthened to J_1. Similarly, the limits $\mathbf{T}_n^A \Rightarrow \mathbf{T}^A$ and $\mathbf{T}_n \to -\lambda \mathbf{T} \circ \lambda e$ cannot both hold in J_1.

Similarly, we cannot have convergence

$$(\mathbf{T}_n^A, \mathbf{D}_n^A) \to (\mathbf{T}^A, \mathbf{T}^A)$$

in $D([0,\infty), \mathbb{R}^2, SJ_1)$ if $\mathbf{S}^v$ has discontinuities, because that would imply that

$$\mathbf{D}_n^A - \mathbf{T}_n^A \to \mathbf{0} \quad \text{in} \quad (D, J_1) \ . \tag{3.36}$$

The limit (3.36) is a contradiction because $\mathbf{S}_n^v \to \mathbf{S}^v$ in (D, J_1) implies that

$$\begin{aligned} J_t(\mathbf{S}_n^v) &\equiv \sup_{0 \le s \le t} \{|\mathbf{S}_n^v(s) - \mathbf{S}_n^v(s-)|\} = J_t(\mathbf{D}_n^A - \mathbf{T}_n^A) \\ &= \sup_{0 \le s \le t} \{c_n^{-1} V_{n,\lfloor ns \rfloor}\} \to J_t(\mathbf{S}^v) \end{aligned}$$

for any t such that $P(t \in Disc(\mathbf{S}^v)) = 0$, and $P(J_t(\mathbf{S}^v) = 0) < 1$ for all sufficiently large t if $\mathbf{S}^v$ fails to have continuous sample paths. ∎

It is natural to choose the measuring units so that the mean service time is 1. Then $\lambda = \mu = \mu_n = 1$ for all n. When $\lambda = 1$, the limit processes $\mathbf{W}$, $\mathbf{L}$, $\mathbf{Q}$ and $\mathbf{Q}^A$ all coincide; they all become $\phi(\mathbf{S})$, the reflection of $\mathbf{S}$. We observed the coincidence of $\mathbf{W}$ and $\mathbf{Q}$ when $\lambda = 1$ in Chapter 6.

The discussion about heavy-traffic scaling in Section 5.5 applies here as well. There are slight differences because (3.11) differs from (4.6) in Section 5.4. If $c_n = n^H$ for $0 < H < 1$ and $\eta < 0$, then just as in (5.10) in Section 5.5, we obtain

$$n = (\zeta/(1-\rho))^{1/(1-H)} \ , \tag{3.37}$$

but now

$$\zeta = -\eta\lambda > 0 \ . \tag{3.38}$$

(Now λ^{-1} plays the role of μ before.)

Remark 9.3.4. *From queue lengths to waiting times.* We conclude this section by mentioning some supplementary material in the Internet Supplement. In Section 5.4 of the Internet Supplement, drawing upon Puhalskii (1994), we show how heavy-traffic limits for workload and waiting-time processes can be obtained directly from associated heavy-traffic limits for arrival, departure and queue-length processes. These results apply the FCLT

for inverse processes with nonlinear centering in Section 13.7. Following Puhalskii (1994), we apply the result to establish a limit for a single-server queue in a central-server model (i.e., a special closed queueing network) as the number of customers in the network increases. In that setting it is not easy to verify the conditions in the earlier limit theorems for waiting times and the workload because the arrival and service processes are state-dependent. That result has also been applied by Mandelbaum, Massey, Reiman and Stolyar (1999).

9.4. Superposition Arrival Processes

In Section 8.3 we established heavy-traffic stochastic-process limits for a fluid queue with input from multiple sources. We now establish analogous heavy-traffic stochastic-process limits for the standard single-server queue with arrivals from multiple sources. We use the inverse map with centering (Sections 13.7 and 13.8) to relate the arrival-time sequences to the arrival counting processes.

With multiple sources, the arrival process in the single-server queue is the superposition of m component arrival processes, i.e.,

$$A(t) \equiv A_1(t) + \cdots + A_m(t), \quad t \geq 0 \ , \tag{4.1}$$

where $\{A_i(t) : t \geq 0\}$ is the i^{th} component arrival counting process with associated arrival times (partial sums)

$$S^u_{i,k} \equiv \inf\{t \geq 0 : A_i(t) \geq k\}, \quad k \geq 0 \ , \tag{4.2}$$

and interarrival times

$$U_{i,k} \equiv S^u_{i,k} - S^u_{i,k-1}, \quad k \geq 1 \ . \tag{4.3}$$

We extend the previous limit theorems in Section 9.3 to this setting by showing how limits for the m partial-sum sequences $\{S^u_{i,k} : k \geq 1\}$, $1 \leq i \leq m$, imply limits for the overall partial-sum sequence $\{S^u_k : k \geq 1\}$, where

$$S^u_k \equiv \inf\{t \geq 0 : A(t) \geq k\}, \quad k \geq 0 \ . \tag{4.4}$$

As in Section 9.3, we consider a sequence of models indexed by n; e.g., let $S^u_{n,i,k}$ be the k^{th} partial sum (arrival time of customer k) in the i^{th} component arrival process of model n. Let the random elements of D be

defined by

$$\begin{aligned}
\mathbf{S}^u_{n,i}(t) &\equiv c_n^{-1}[S^u_{n,i,\lfloor nt \rfloor} - \lambda^{-1}_{n,i} nt] \\
\mathbf{A}_{n,i}(t) &\equiv c_n^{-1}[A_{n,i}(nt) - \lambda_{n,i} nt] \\
\mathbf{S}^u_{n}(t) &\equiv c_n^{-1}[S^u_{n,\lfloor nt \rfloor} - \lambda^{-1}_{n} nt] \\
\mathbf{A}_{n}(t) &\equiv c_n^{-1}[A_{n}(nt) - \lambda_{n} nt], \quad t \geq 0 \ , \qquad (4.5)
\end{aligned}$$

for $n \geq 1$. The M_1 topology plays an important role when the limit processes can have discontinuous sample paths.

Theorem 9.4.1. (FCLT for superposition arrival processes) *Suppose that*

$$(\mathbf{S}^u_{n,1}, \ldots, \mathbf{S}^u_{n,m}) \Rightarrow (\mathbf{S}^u_1, \ldots, \mathbf{S}^u_m) \quad \textit{in} \quad (D^m, WM_1) \ , \qquad (4.6)$$

where $\mathbf{S}^u_{n,i}$ *is defined in* (4.5),

$$P(Disc(\mathbf{S}^u_i \circ \lambda_i \mathbf{e}) \cap Disc(\mathbf{S}^u_j \circ \lambda_j \mathbf{e}) = \phi) = 1 \qquad (4.7)$$

for all i, j *with* $1 \leq i, j \leq m$ *and* $i \neq j$, *and*

$$P(\mathbf{S}^u_i(0) = 0) = 1, \quad 1 \leq i \leq m \ . \qquad (4.8)$$

If, in addition $c_n \to \infty$, $n/c_n \to \infty$ *and* $\lambda_{n,i} \to \lambda_i$, $0 < \lambda_i < \infty$, *for* $1 \leq i \leq m$, *then*

$$\begin{aligned}
&(\mathbf{S}^u_{n,1}, \ldots, \mathbf{S}^u_{n,m}, \mathbf{A}_{n,1}, \ldots, \mathbf{A}_{n,m}, \mathbf{A}_n, \mathbf{S}^u_n) \\
&\quad \Rightarrow (\mathbf{S}^u_1, \ldots, \mathbf{S}^u_m, \mathbf{A}_1, \ldots, \mathbf{A}_m, \mathbf{A}, \mathbf{S}^u) \qquad (4.9)
\end{aligned}$$

in (D^{2m+2}, WM_1), *where*

$$\lambda_n \equiv \lambda_{n,1} + \cdots + \lambda_{n,m} \ ,$$

$$\begin{aligned}
\mathbf{A}_i &\equiv -\lambda_i \mathbf{S}^u_i \circ \lambda_i \mathbf{e}, \quad \mathbf{A} \equiv \mathbf{A}_1 + \cdots + \mathbf{A}_m \\
\mathbf{S}^u &\equiv -\lambda^{-1} \mathbf{A} \circ \lambda^{-1} \mathbf{e} = \sum_{i=1}^m \gamma_i \mathbf{S}^u_i \circ \gamma_i \mathbf{e} \qquad (4.10)
\end{aligned}$$

for

$$\lambda \equiv \lambda_1 + \cdots + \lambda_m \quad \textit{and} \quad \gamma_i \equiv \lambda_i / \lambda, \quad 1 \leq i \leq m \ . \qquad (4.11)$$

Proof. We apply the Skorohod representation theorem, Theorem 3.2.2, to replace the convergence in distribution in (4.6) by convergence w.p.1 for special versions. We then apply the convergence-preservation results for the inverse map with centering, as applied to counting functions, in Corollary 13.8.1 to obtain, first, the limits for $\mathbf{A}_{n,i}$ from the limits for $\mathbf{S}^u_{n,i}$ and, second, the limit for $\mathbf{S}^u_n$ from the limit for $\mathbf{A}_n$. We use addition with condition (4.7) to obtain the convergence of $\mathbf{A}_n$ from the convergence of $(\mathbf{A}_{n,1}, \ldots, \mathbf{A}_{n,m})$. ∎

Remark 9.4.1. *The case of IID Lévy processes.* Suppose that the limit processes $\mathbf{S}^u_1, \ldots, \mathbf{S}^u_m$ in Theorem 9.4.1 are IID Lévy processes. Then $\gamma_i = m^{-1}$,

$$\mathbf{A} \stackrel{\mathrm{d}}{=} \mathbf{A}_1 \circ m\mathbf{e} \quad \text{and} \quad \sum_{i=1}^{m} \mathbf{S}^u_i \stackrel{\mathrm{d}}{=} \mathbf{S}^u_1 \circ m\mathbf{e} \ ,$$

so that

$$\mathbf{S}^u \stackrel{\mathrm{d}}{=} m^{-1}\mathbf{S}^u_1 \ .$$

In this case, $\mathbf{A}$ and $\mathbf{S}$ differ from $\mathbf{A}_1$ and $\mathbf{S}^u_1$ only by the deterministic scale factor m.

We can remove the deterministic scale factor by rescaling to make the overall arrival rate independent of m. We can do that for any given m by replacing $A_i(t)$ by $A_i(t/m)$ for $t \geq 0$ or, equivalently, by replacing $S^u_{i,k}$ by $m\, S^u_{i,k}$ for $k \geq 0$. If we make that scale change at the outset, then the limit processes $\mathbf{A}$ and $\mathbf{S}^u$ become independent of m. However, we cannot draw that conclusion if the limit processes $\mathbf{S}^u_1, \ldots, \mathbf{S}^u_m$ are not Lévy processes. For further discussion, see Section 5.6 and Remarks 10.2.2 and 10.2.4. ∎

We can combine Theorems 9.3.3, 9.3.4 and 9.4.1 to obtain a heavy-traffic limit for queues with superposition arrival processes.

Theorem 9.4.2. (heavy-traffic limit for a queue with a superposition arrival process) *Suppose that*

$$(\mathbf{S}^u_{n,1}, \ldots, \mathbf{S}^u_{n,m}, \mathbf{S}^v_n) \Rightarrow (\mathbf{S}^u_1, \ldots, \mathbf{S}^u_m, \mathbf{S}^v) \quad \textit{in} \quad (D^{m+1}, WM_1) \tag{4.12}$$

for $\mathbf{S}^u_{n,i}$ *in* (4.5) *and* $\mathbf{S}^v_n$ *in* (3.1), *where*

$$P(Disc(\mathbf{S}^u_i \circ \gamma_i\mathbf{e}) \cap Disc(\mathbf{S}^u_j \circ \gamma_j\mathbf{e}) = \phi) = 1 \tag{4.13}$$

and

$$P(Disc(\mathbf{S}^u_i \circ \gamma_i\mathbf{e}) \cap Disc(\mathbf{S}^v) = \phi) = 1 \tag{4.14}$$

for all i, j *with* $1 \leq i, j \leq m$, $i \neq j$ *and* γ_i *in* (4.11). *Suppose that, for* $1 \leq i \leq m$,

$$P(\mathbf{S}_i^u(0) = 0) = P(\mathbf{S}^v(0) = 0) = 1 \ , \tag{4.15}$$

$c_n \to \infty$, $n/c_n \to \infty$, $\lambda_{n,i}^{-1} \to \lambda_i^{-1}$, $0 < \lambda_i^{-1} < \infty$, *and*

$$\eta_n \equiv n(\mu_n^{-1} - \lambda_n^{-1})c_n \to \eta \quad as \quad n \to \infty \tag{4.16}$$

for λ_n *in* (4.10). *Then the conditions and conclusions of Theorems* 9.3.3 *and* 9.3.4 *hold with* $\mathbf{S}^u$ *and* $\mathbf{A}$ *in* (4.10) *and* λ *in* (4.11).

Proof. As usual, start by applying the Skorohod representation theorem to replace convergence in distribution by convergence w.p.1, without introducing special notation for the special versions. Then conditions (4.12)–(4.15) plus Theorem 9.4.1 imply that conditions (3.9) and (3.10) in Theorem 9.3.3 and condition (3.19) in Theorem 9.3.4 hold. Thus the conditions of Theorems 9.3.3 and 9.3.4 hold with $\mathbf{S}^u$ and $\mathbf{A}$ in (4.10) and λ in (4.11). ∎

We now show what Theorem 9.4.2 yields in the standard light-tailed weak-dependent case. The following results closely parallels Theorem 8.4.1.

Corollary 9.4.1. (the Brownian case) *Suppose that the conditions of Theorem* 9.4.2 *hold with* $c_n = \sqrt{n}$, $\mathbf{S}_i^u = \sigma_{u,i}\mathbf{B}_i^u$, $1 \leq i \leq m$, *and* $\mathbf{S}^v = \sigma_v \mathbf{B}^v$, *where* $\mathbf{B}_1^u, \ldots, \mathbf{B}_m^u, \mathbf{B}^v$ *are* $m+1$ *IID standard Brownian motions. Then the conclusions of Theorem* 9.4.2 *hold with*

$$\mathbf{S}^u \stackrel{\mathrm{d}}{=} \sigma_u \mathbf{B} \tag{4.17}$$

for

$$\sigma_u^2 = \sum_{i=1}^{m} \gamma_i^3 \sigma_{u,i}^2 \tag{4.18}$$

and

$$\mathbf{S} \stackrel{\mathrm{d}}{=} \sigma_S \mathbf{B} + \eta \mathbf{e} \tag{4.19}$$

for η *in* (4.16),

$$\sigma_S^2 = \sigma_u^2 + \sigma_v^2 \ , \tag{4.20}$$

and $\mathbf{B}$ *being a standard Brownian motion.*

A corresponding corollary is easy to establish in the heavy-tailed case, when the limits are scaled versions of independent stable Lévy motions. For the IID case, using essentially a single model, we apply Theorem 4.5.3. Since the random variables $U_{n,i,k}$ and $V_{n,k}$ are nonnegative, we get totally skewed stable Lévy motion limits (with $\beta = 1$) for $\mathbf{S}_i^u$ and $\mathbf{S}^v$.

Corollary 9.4.2. (the stable-Lévy-motion case) *Suppose that the conditions of Theorem* 9.4.2 *hold with the limit processes* $\mathbf{S}_i^u$, $1 \le i \le m$, *and* $\mathbf{S}^v$ *being mutually independent stable Lévy motions with index* α, $1 < \alpha < 2$, *where*

$$\mathbf{S}_i^u(1) \stackrel{\mathrm{d}}{=} \sigma_{u,i} S_\alpha(1,1,0), \quad 1 \le i \le m \ , \tag{4.21}$$

and

$$\mathbf{S}^v(1) \stackrel{\mathrm{d}}{=} \sigma_v S_\alpha(1,1,0) \ . \tag{4.22}$$

Then the conclusions of Theorem 9.4.2 *hold with* $\mathbf{S}^u$ *and* $\mathbf{S}$ *being stable Lévy motions with index* α, *where*

$$\mathbf{S}^u(1) \stackrel{\mathrm{d}}{=} \sigma_u S_\alpha(1,1,0) \tag{4.23}$$

for

$$\sigma_u = \left(\sum_{i=1}^m \gamma_i^{\alpha+1} \right)^{1/\alpha}$$

and

$$\mathbf{S}(1) \stackrel{\mathrm{d}}{=} S_\alpha(\sigma, \beta, 0) \ , \tag{4.24}$$

where

$$\sigma = (\sigma_v^\alpha + \sigma_u^\alpha)^{1/\alpha}$$

and

$$\beta = \frac{\sigma_v^\alpha - \sigma_u^\alpha}{\sigma_v^\alpha + \alpha_u^\alpha} \ .$$

Proof. Again we apply Theorem 9.4.2. We obtain (4.23) and (4.24) by applying the basic scaling relations in (5.7)–(5.11) of Section 4.5. ∎

9.5. Split Processes

In this section we obtain a FCLT for counting processes that are split from other counting processes. For example, the original counting process might be a departure process, and each of these departures may be routed to one of several other queues. We then want to consider the arrival counting processes at these other queues. We also allow new points to be created in these split arrival processes. (Events in the original process may trigger or cause one or more events of different kinds. In manufacturing there may be batching and unbatching. In communication networks there may be multicasting; the same packet received may be simultaneously sent out on several outgoing links.)

Let $\tilde{A}(t)$ count the number of points in the original process in the time interval $[0,t]$. Let $X_{i,j}$ be the number of points assigned to split process i at the epoch of the j^{th} point in the original arrival process $\tilde{A}$. With the standard splitting, for each j, $X_{i,j} = 1$ for some i and $X_{i,j} = 0$ for all other i, but we allow other possibilities.

Under the assumptions above, the number of points in the i^{th} split counting process in the time interval $[0,t]$ is

$$A_i(t) \equiv \sum_{j=1}^{\tilde{A}(t)} X_{i,j}, \quad t \geq 0 \ . \tag{5.1}$$

Now we assume that we have processes as above for each n, i.e., $\{\tilde{A}_n(t) : t \geq 0\}$, $\{X_{n,i,j} : i \geq 1\}$ and $\{A_{n,i}(t) : t \geq 0\}$. We form associated random elements of $D \equiv D([0,\infty), \mathbb{R})$ by setting

$$\begin{aligned} \tilde{\mathbf{A}}_n(t) &\equiv c_n^{-1}[\tilde{A}_n(nt) - \lambda_n nt] \\ \mathbf{S}_{n,i}(t) &\equiv c_n^{-1}\left[\sum_{j=1}^{\lfloor nt \rfloor} X_{n,i,j} - p_{n,i} nt\right] \\ \mathbf{A}_{n,i}(t) &\equiv c_n^{-1}[A_{n,i}(nt) - \lambda_n p_{n,i} nt], \quad t \geq 0 \,, \end{aligned} \tag{5.2}$$

where λ_n is a positive scalar and $p_n \equiv (p_{n,1}, \ldots, p_{n,m})$ is an element of $\mathbb{R}^m$ with nonnegative components.

We can apply Corollary 13.3.2 to establish a FCLT for the vector-valued split processes $\mathbf{A}_n \equiv (\mathbf{A}_{n,1}, \ldots, \mathbf{A}_{n,m})$ in D^m. Let $\mathbf{S}_n \equiv (\mathbf{S}_{n,1}, \ldots, \mathbf{S}_{n,m})$.

Theorem 9.5.1. (FCLT for split processes) *Suppose that*

$$(\tilde{\mathbf{A}}_n, \mathbf{S}_n) \Rightarrow (\tilde{\mathbf{A}}, \mathbf{S}) \quad \textit{in} \quad D^{1+m} \tag{5.3}$$

with the topology WJ_1 or WM_1. Also suppose that $c_n \to \infty$, $n/c_n \to \infty$, $\lambda_n \to \lambda$, $p_n \to p$, where $p_i > 0$ for each i, and almost surely $\tilde{\mathbf{A}}$ and $\mathbf{S}_i \circ \lambda\mathbf{e}$ have no common discontinuities of opposite sign for $1 \leq i \leq m$. Then

$$\mathbf{A}_n \Rightarrow \mathbf{A} \quad \textit{in} \quad D^m \tag{5.4}$$

with the same topology, where

$$\mathbf{A}_i \equiv p_i \tilde{\mathbf{A}} + \mathbf{S}_i \circ \lambda\mathbf{e} \ . \tag{5.5}$$

Proof. Since $A_i(t)$ in (5.1) is a random sum, we can apply the continuous mapping theorem, 3.4.3, with composition and addition. Specifically, we apply Corollary 13.3.2 after noting that

$$\mathbf{A}_{n,i} = \mathbf{S}_{n,i} \circ \hat{\mathbf{A}}_{n,i} + p_{n,i}\tilde{\mathbf{A}}_{n,i} \ ,$$

where $\hat{\mathbf{A}}_{n,i} \equiv n^{-1}A_{n,i}(nt),, \ t \geq 0$. ∎

If $c_n/\sqrt{n} \to \infty$, it will often happen that one of the limit processes $\tilde{\mathbf{A}}$ or $\mathbf{S}_i$ will be the zero function. If the burstiness in $\tilde{\mathbf{A}}$ dominates, so that $\mathbf{S}_i = 0\mathbf{e}$, then the limit in (5.4) becomes $p_i\tilde{\mathbf{A}}$, a deterministic-scalar multiple of the limit process $\tilde{\mathbf{A}}$. On the other hand, if the burstiness in $\mathbf{S}_i$ dominates, so that $\tilde{\mathbf{A}} = 0\mathbf{e}$, then the limit in (5.4) becomes $\mathbf{S}_i \circ \lambda\mathbf{e}$, a deterministic time change of the limit process $\mathbf{S}_i$.

It is instructive to contrast various routing methods. Variants of the round robin discipline approximate deterministic routing, in which every $(1/p_i)^{\text{th}}$ arrival from $\tilde{A}$ is assigned to A_i. With any reasonable approximation to round robin, we obtain $\mathbf{S}_i = 0\mathbf{e}$.

In contrast, with IID splittings, $\sum_{j=1}^{k} X_{n,i,j}$ has a binomial distribution for each n, i and k, so that $\mathbf{S}_{n,i} \Rightarrow \mathbf{S}_i$, where $c_n = \sqrt{n}$, $\mathbf{S}_i \stackrel{\text{d}}{=} \sigma_i\mathbf{B}$ with $\mathbf{B}$ standard Brownian motion and $\sigma_i^2 = p_i(1-p_i)$. Then $\mathbf{S}$ is a zero-drift Brownian motion with covariance matrix $\Sigma \equiv (\sigma_{S,i,j}^2)$, where $\sigma_{S,i,i}^2 = p_i(1-p_i)$ and $\sigma_{S,i,j}^2 = -p_ip_j$ for $i \neq j$. We thus see that IID splitting produces greater variability in the split arrival processes than round robin, and thus produces greater congestion in subsequent queues. Moreover, with the heavy-traffic stochastic-process limits, we can quantify the difference.

9.6. Brownian Approximations

In this section we continue the discussion begun in Section 5.7 of Brownian limits that occur in the light-tailed weak-dependent case and the associated Brownian (or RBM) approximations that stem from them.

In the standard light-tailed weak-dependent case, the conditions of Theorems 9.3.1–9.3.4 hold with space scaling by $c_n = \sqrt{n}$ and limits

$$\mathbf{S} \stackrel{\text{d}}{=} \sigma_S\mathbf{B} + \eta\mathbf{e} \ , \tag{6.1}$$

where $\mathbf{B}$ is standard Brownian motion and η is obtained from the limit (3.11). Just as in Section 5.7, we can obtain such limits by considering essentially a single model. Here the single model is based on a single sequence of interarrival times and service times $\{(U_k, V_k) : k \geq 1\}$. Let the associated

partial sums be $S_k^u \equiv U_1 + \cdots U_k$ and $S_k^v \equiv V_1 + \cdots + V_k$, $k \geq 1$. We then construct the sequences $\{(U_{n,k}, V_{n,k}) : k \geq 1\}$ for a sequence of models indexed by n by scaling the service times, i.e., by letting

$$U_{n,k} = U_k \quad \text{and} \quad V_{n,k} \equiv \rho_n V_k \ . \tag{6.2}$$

Then, in the setting of Section 9.3, $\lambda_n = \lambda$ and $\mu_n^{-1} = \lambda^{-1}\rho_n$ for all n. Then condition (3.11) becomes

$$\sqrt{n}(1 - \rho_n) \to \zeta \equiv -\eta\lambda > 0 \quad \text{as} \quad n \to \infty \ . \tag{6.3}$$

The required FCLT for (S_n^u, S_n^v) in condition (3.9) then follows from Donsker's theorem in Section 4.3 or one of its generalizations for dependent sequences in Section 4.4, applied to the partial sums of the single sequences $\{(U_k, V_k)\}$, under the assumptions there.

As in Section 5.5, it is natural to index the family of queueing systems by the traffic intensity ρ, where $\rho \uparrow 1$. Then, focusing on the waiting-time and queue-length processes and replacing n by $\zeta^2(1-\rho)^{-2}$ for ζ in (3.38) and (6.3), we have the Brownian approximations

$$W_{\rho,k} \approx \lambda\sigma_S^2(1-\rho)^{-1}\mathbf{R}(\lambda^{-2}\sigma_S^{-2}(1-\rho)^2; -1, 1, 0) \tag{6.4}$$

and

$$Q_\rho(t) \approx \lambda^2\sigma_S^2(1-\rho)^{-1}\mathbf{R}(\lambda^{-1}\sigma_S^{-2}(1-\rho)^2; -1, 1, 0) \ , \tag{6.5}$$

where $\{\mathbf{R}(t; -1, 1, 0) : t \geq 0\}$ is canonical RBM. The Brownian approximation in (6.4) is the same as the Brownian approximation in (7.8) in Section 5.7 with λ^{-1} replacing μ. Approximation (6.5) follows from (6.4) because $\mathbf{Q} = \lambda\mathbf{W} \circ \lambda\mathbf{e}$; see (3.20).

9.6.1. Variability Parameters

For the GI/GI/1 queue, where the basic sequences $\{U_k\}$ and $\{V_k\}$ are independent sequences of IID random variables, the heavy-traffic variance constant is

$$\sigma_S^2 = \sigma_u^2 + \sigma_v^2 \ , \tag{6.6}$$

where

$$\sigma_u^2 \equiv Var\, U_1 \quad \text{and} \quad \sigma_v^2 \equiv Var\, V_1 \ .$$

For better understanding, it is helpful to replace the variances by dimensionless variability parameters: It is convenient to use the *squared coefficients of variation* (SCVs), defined by

$$c_u^2 \equiv \frac{Var\, U_1}{(EU_1)^2} \quad \text{and} \quad c_v^2 \equiv \frac{Var\, V_1}{(EV_1)^2} \ . \tag{6.7}$$

Combining (6.5)–(6.7), we have

$$\sigma_S^2 = \frac{c_u^2 + c_v^2}{\lambda^2} \equiv \frac{c_{HT}^2}{\lambda^2} \ , \tag{6.8}$$

where c_{HT}^2 is the dimensionless overall variability parameter.

For the more general G/G/1 queue, in which $\{(U_k, V_k) : k \geq 1\}$ is a stationary sequence, we must include covariances. In particular,

$$\sigma_S^2 = c_{HT}^2/\lambda^2, \tag{6.9}$$

where

$$c_{HT}^2 = c_U^2 + c_V^2 - 2c_{U,V}^2 \ , \tag{6.10}$$

with

$$\begin{aligned} c_U^2 &\equiv \lim_{k\to\infty} k^{-1} \frac{Var\, S_k^u}{(EU_1)^2} \equiv \lim_{k\to\infty} k^{-1} \sum_{j=1}^{k} (k-j) \frac{Cov(U_1, U_j)}{(EU_1)^2} \ , \\ c_V^2 &\equiv \lim_{k\to\infty} k^{-1} \frac{Var\, kS_k^v}{(EV_1)^2} \equiv \lim_{k\to\infty} k^{-1} \sum_{j=1}^{k} (k-j) \frac{Cov(V_1, V_0)}{(EV_1)^2} \ , \\ c_{U,V}^2 &\equiv \lim_{k\to\infty} k^{-1} \frac{Cov(S_k^u, S_k^v)}{(EU_1)(EV_1)} \equiv \lim_{k\to\infty} k^{-1} \sum_{j=1}^{k} (k-j) \frac{Cov(U_1, V_j)}{(EU_1)(EV_1)} \end{aligned} \tag{6.11}$$

We call c_U^2, c_V^2 and $c_{U,V}^2$ in (6.11) the *asymptotic variability parameters* for the arrival and service processes.

We can combine (6.4), (6.5) and (6.9) to obtain general Brownian approximations in terms of the dimensionless variability parameter c_{HT}^2:

$$\begin{aligned} W_{\rho,k} &\approx \lambda^{-1} c_{HT}^2 (1-\rho)^{-1} \mathbf{R}(c_{HT}^{-2}(1-\rho)^2 k; -1, 1, 0) \\ Q_\rho(t) &\approx c_{HT}^2 (1-\rho)^{-1} \mathbf{R}(c_{HT}^{-2} \lambda (1-\rho)^2 t; -1, 1, 0) \ . \end{aligned} \tag{6.12}$$

For example, as a consequence, the approximation for the mean steady-state waiting time is

$$EW_{\rho,\infty} \approx \lambda^{-1} c_{HT}^2 / 2(1-\rho) \ . \tag{6.13}$$

(Recall that the mean service time in model ρ is ρ here.)

The dimensionless variability parameter c_{HT}^2 helps to understand the heavy-traffic limits for queues with superposition arrival processes. If the arrival process is the superposition of m IID component arrival processes, then c_{HT}^2 is independent of the number m of processes. (See Remark 9.4.1.)

For the GI/GI/1 model, $c_U^2 = c_u^2$, $c_V^2 = c_v^2$ and $c_{U,V}^2 = 0$. However, in many more general G/G/1 applications, these relations do not nearly hold. For example, that usually is the case with superposition arrival processes arising in models of statistical multiplexing in communication networks.

Example 9.6.1. *A packet network example.* In a detailed simulation of a packet network link (specifically, an X.25 link) with 25 independent sources, Fendick, Saksena and Whitt (1989) found that

$$c_u^2 \approx 1.89, \quad c_v^2 \approx 1.06 \quad \text{and} \quad c_{u,v}^2 \approx 0.03 \ , \tag{6.14}$$

where c_u^2 and c_v^2 are in (6.7) and

$$c_{u,v}^2 \equiv \frac{Cov(U_1, V_1)}{(EU_1)(EV_1)} \ . \tag{6.15}$$

In contrast, they found that

$$c_U^2 \approx 17.6, \quad c_V^2 \approx 35.1 \quad \text{and} \quad c_{U,V}^2 \approx -6.7 \ . \tag{6.16}$$

The differences between (6.16) and (6.14) show that there are significant correlations: (i) among successive interarrival times, (ii) among successive service times and (iii) between interarrival times and service times. The dependence among service times and between service times and interarrival times occur because of bursty arrivals from multiple sources with different mean service times (due to different packet lengths).

Note that the variability parameter c_{HT}^2 based on (6.10) and (6.16) is very different from the one based on (6.7), (6.8), (6.14) and (6.16). The variability parameter based on (6.10) and (6.16) is

$$c_{HT}^2 \approx 17.6 + 35.1 - 2(6.7) = 66.1 \ . \tag{6.17}$$

If, instead, we acted as if we had a GI/GI/1 queue and used (6.7), (6.8) and (6.14), we would obtain $c_{HT}^2 = 2.79$.

Moreover, under moderate to heavy loads, the average steady-state queue lengths in the simulation experiments were consistent with formulas (6.10) and (6.16) using the variability parameter c_{HT}^2 in (6.17). However, under lighter loads there were significant differences between the observed average queue lengths and the heavy-traffic approximations, which motivate alternative parametric approximations that we discuss in Section 9.9 below.

This simulation experiment illustrates that correlations can be, not only an important part of the relevant variability, but the dominant part; in this example,

$$c_U^2 + c_V^2 - 2c_{U,V}^2 >> c_u^2 + c_v^2 \ . \tag{6.18}$$

Moreover, in this example the lag-k correlations, defined by

$$\begin{aligned} c_{u,k}^2 &= \frac{Cov(U_1, U_{1+k})}{(EU_1)^2}, \quad c_{v,k}^2 \equiv \frac{Cov(V_1, V_{1+k})}{(EV_1)^2} \\ c_{u,v,k}^2 &= \frac{Cov(U_1, V_{1+k})}{(EU_1, EV_1)} , \end{aligned} \tag{6.19}$$

were individually small for all k. The values in (6.16) were substantially larger than those in (6.14) because of the cumulative effect of many small correlations (over all k). See Albin (1982) for similar experiments. ∎

9.6.2. Models with More Structure

The heavy-traffic Brownian approximation is appealing because it is often not difficult to compute the variability parameter c_{HT}^2 in (6.10) for models. Indeed, in Section 4.4 we indicated that it is often possible to compute the normalization constant in a CLT involving dependent summands. There the specific formulas and algorithms depended on Markov structure. Now we illustrate by considering a model from Fendick, Saksena and Whitt (1989, 1991) that has more structure.

We consider a multi-class batch-renewal-process model that might serve as a model for a packet arrival process in a communication network. In that context, a customer class can be thought of as a particular kind of traffic such as data, video or fax. As an approximation, we assume that all packets (customers) in the same batch (message, burst or flow) arrive at the same instant. We discuss generalizations afterwards in Remark 9.6.1. For this model, we show how to determine the variability parameters c_U^2, c_V^2 and $c_{U,V}^2$.

We assume that the arrival process of k customer classes come as mutually independent batch-renewal processes. For class i, batches arrive according to a renewal process with rate λp_i where the interrenewal-time cdf has SCV $c_{u,i}^2$; the successive batch sizes are IID with mean m_i and SCV $c_{b,i}^2$; the packet service times are IID with mean τ_i and SCV $c_{v,i}^2$. (We assume that $p_1 + \cdots p_k = 1$, so that the total arrival rate of batches is λ. The total arrival rate of customers (packets) is thus $\bar{\lambda} \equiv \lambda m_B$ where

$$m_B = \sum_{i=1}^{k} p_i m_i \ . \tag{6.20}$$

Let q_i be the probability that an arbitrary packet belongs to class i, i.e.,

$$q_i \equiv p_i m_i / \sum_{j=1}^{k} p_j m_j \ . \tag{6.21}$$

Let

$$\tau \equiv \frac{\sum_{i=1}^{k} p_i m_i \tau_i}{\sum_{i=1}^{k} p_i m_i} \quad \text{and} \quad r_i \equiv \frac{\tau_i}{\tau} \ . \tag{6.22}$$

We do not describe the full distributions of intervals between batches, batch sizes and service times, because the heavy-traffic limit does not depend on that extra detail. The model can be denoted by $\sum(GI^{B_i}/GI_i)/1$, since the service times are associated with the arrivals.

Let $c_{U,i}^2$ be the heavy-traffic variability parameter for the class-i arrival process alone.

Theorem 9.6.1. (Heavy-traffic limit for the $\sum(GI^{B_i}/GI)/1$ model) *For the single-server queue with multi-class batch-renewal input above, the conditions of Theorems* 9.3.1 *and* 9.3.3 *hold with* $c_n = n^{1/2}$ *and* $(\mathbf{S}^u, \mathbf{S}^v)$ *being two-dimensional zero-drift Brownian motion, supporting the approximations in* (6.12) *with*

$$\begin{aligned} c_U^2 &= \sum_{i=1}^{k} q_i c_{U,i}^2 \ , \\ c_V^2 &= \sum_{i=1}^{k} q_i [r_i^2 c_{v,i}^2 + (r_i - 1)^2 c_{U,i}^2] \ , \\ c_{U,V}^2 &= \sum_{i=1}^{k} q_i (1 - r_i) c_{U,i}^2 \ , \\ c_{U,i}^2 &= m_i (c_{b,i}^2 + c_{u,i}^2) \ . \end{aligned} \tag{6.23}$$

Proof. We only give a quick sketch. The independence assumptions allow us to obtain FCLTs for the partial sums of the batch interarrival times, the batch sizes and the service times. Given that initial FCLT, we can apply the Skorohod representation theorem to replace the convergence in distributions by convergence w.p.1 for special versions. Then note that the packet arrival process can be represented as a random sum: the number of packet arrivals in $[0, t]$ is the sum of the IID batches up to the number of batches to arrive in $[0, t]$. Hence we can apply Corollary 13.3.2 for random

sums. The overall packet counting process is the sum of the k independent class packet counting processes. The partial sums of the interarrival times can be treated as the inverses of the counting processes. We thus obtain the limits for all arrival processes and $\mathbf{S}_n^u$, and the variability parameters $c_{U,i}^2$ and c_U^2 in (6.23). We treat the total input of work by adding over the classes, with the total input of work for each class being a random sum of the IID service times up to the number of packet arrivals. Hence we can apply Corollary 13.3.2 for random sums again. From the total input of work, we can directly obtain the limit for the workload by applying the reflection map. From the limit for the total input of work, we can also obtain a limit for the service times presented in order of their arrival to the queue. (This is perhaps the only tricky step.) In general, we have

$$C_n(S_{n,k}^u-) \le S_{n,k}^v \le C_n(S_{n,k}^u) \quad \text{for all} \quad n, k\,, \tag{6.24}$$

where $S_{n,k}^v$ is the k^{th} partial sum of the service times associated with successive arrivals in model n. We first obtain the limit for $C_n(S_{n,k}^u)$ by applying the random-sum result in Corollary 13.3.2 once more. Since the limit process has continuous sample paths, from (6.24) we can conclude that $\mathbf{S}_n^v$ has the same limit; see Corollary 12.11.6. Given the limit for $(\mathbf{S}_n^u, \mathbf{S}_n^v)$, we can apply Theorems 9.3.3 and 9.3.4. ∎

It is helpful to further interpret the asymptotic variability parameters in (6.23). Note that c_U^2 is a convex combination of $c_{U,i}^2$ weighted by q_i in (6.21), where $q_1 + \cdots + q_k = 1$. Note that $c_{U,i}^2$ is directly proportional to the mean batch size m_i. Note that c_V^2 and $c_{U,V}^2$ also can be represented as weighted sums of $c_{V,i}^2$ and $c_{U,V,i}^2$, where

$$\begin{aligned} c_{V,i}^2 &\equiv r_i^2 c_{v,i}^2 + (r_i - 1)^2 c_{U,i}^2 \\ c_{U,V,i}^2 &\equiv (1 - r_i) c_{U,i}^2 \end{aligned} \tag{6.25}$$

The class-i asymptotic service variability parameter $c_{U,i}^2$ and the class-i covariance asymptotic parameter $c_{U,V,i}^2$ depend upon the non-class-i processes only via the parameter $r_i \equiv \tau_i/\tau \equiv \tau_i$ in (6.22). Note that r_i is large (small) when class-i service times are larger (smaller) than usual. Note that $c_{V,i}^2$ has the component $r_i^2 c_{v,i}^2$ that is directly proportional to r_i^2 and $c_{v,i}^2$.

Remark 9.6.1. *Extra dependence.* In Theorem 9.6.1 we assumed that the basic model sequences are independent sequences of IID random variables. Using Section 4.4 that can be greatly relaxed. In the spirit of Section 9.3, we could have started with a joint FCLT.

For the model in Theorem 9.6.1, we let all arrivals in a batch arrive at the same instant. We could instead allow the arrivals from each batch to arrive in some arbitrary manner in the interval between that batch arrival and the next. It is significant that Theorem 9.6.1 is unchanged under that modification. However, both the original model and the generalization above implicitly assume that each batch size is independent of the interval between batch arrivals. That clearly is not realistic in many scenarios, e.g., for packet queues, where larger batch sizes usually entail longer intervals between batch arrivals. It is not difficult to create models that represent this feature. In particular, let $\{(B_n^i, L_n^i) : n \geq 1\}$ be the sequence of successive pairs of successive batch sizes and interval length between successive batch arrivals for class i. Assume that successive pairs are IID, but allow B_n^i and L_n^i to be dependent for each n. As above, let m_i and $c_{b,i}^2$ be the two parameters for B_n^i and let $(\lambda p_i)^{-1}$ and $c_{r,i}^2$ be the two parameters for L_n^i. Let $\gamma_{b,r,i}$ be the correlation between B_n^i and L_n^i. Then Theorem 9.6.1 remains valid with $c_{U,i}^2$ in (6.23) replaced by

$$c_{U,i}^2 = m_i(c_{b,i}^2 + c_{r,i}^2 - 2\gamma_{r,b,i}c_{b,i}c_{r,i})\,. \quad \blacksquare \tag{6.26}$$

The multi-class batch-renewal-process model above illustrates that the asymptotic variability parameters c_U^2, c_V^2 and $c_{U,V}^2$ appearing in the expression for c_{HT}^2 in (6.10) can often be computed for quite rich and complex models. We conclude this section by illustrating this feature again for point processes in a random environment, such as the Markov-modulated Poisson process (MMPP).

Example 9.6.2. *Point processes in random environments.* In this example we suppose that the arrival process can be represented as a counting process in a random environment, such as

$$A(t) = X(Y(t)), \quad t \geq 0\ , \tag{6.27}$$

where

$$(\mathbf{X}_n, \mathbf{Y}_n) \Rightarrow (\mathbf{B}_1, \mathbf{B}_2) \quad \text{in} \quad D^2 \tag{6.28}$$

with $(\mathbf{B}_1, \mathbf{B}_2)$ being two-dimensional Brownian motion and

$$(\mathbf{X}_n, \mathbf{Y}_n)(t) \equiv n^{-1/2}(X(nt) - xnt, Y(nt) - ynt), \quad t \geq 0\,. \tag{6.29}$$

For example, an MMPP satisfies (6.27)–(6.29) where X is a homogeneous Poisson process and Y is a function of an irreducible finite-state continuous-time Markov chain (CTMC). Indeed, the representation (6.27) was already

exploited for the cumulative-input processes of the fluid queue in (2.6) in Chapter 8.

Given (6.27), we can obtain the required FCLT for A from an established FCLT for (X, Y) in (6.28) by applying Corollary 13.3.2. Without loss of generality (by deterministically scaling X and Y in (6.27)), we can obtain (6.27) with X, Y and A all being rate-1 processes, i.e., for $x = y = 1$ in (6.29). Then the FCLT for $\mathbf{A}_n$ yields the dimensionless asymptotic variability parameter

$$c_A^2 = c_X^2 + c_Y^2 \ .$$

For example, if A is a rate-1 MMPP and X is a rate-1 Poisson process, then

$$c_A^2 = 1 + c_X^2 \ ,$$

where c_X^2 is the asymptotic variability parameter of a function of a stationary CTMC having mean 1, which is given in Theorem 2.3.4 in the Internet Supplement. ∎

9.7. Very Heavy Tails

When the interarrival times and service times come from independent sequences of IID random variables with heavy-tailed distributions, we obtain heavy-traffic stochastic-process limits with reflected-stable-Lévy-motion (RSLM) limit processes from Sections 4.5 and 9.3, the same way we obtained heavy-traffic stochastic-process limits with RSLM limit processes for fluid queues in Section 8.5 from Sections 4.5, 5.4 and 8.3. For the most part, the story has already been told in Section 8.5. Hence, now we will only discuss the case of very heavy tails, arising when the random variables have infinite mean. See Resnick and Rootzén (2000) for related results.

Specifically, as in (5.26) in Section 4.5, we assume that the service-time distribution has a power tail, satisfying

$$P(V_1 > x) \sim Ax^{-\alpha} \quad \text{as} \quad x \to \infty \tag{7.1}$$

for positive constants α and A with $0 < \alpha < 1$.

We note that $\alpha = 1$ is a critical boundary point, because if (7.1) holds for $\alpha > 1$, then the service-time distribution has a finite mean, which implies that the waiting-time process has a proper steady-state distribution. However, if (7.1) holds for $\alpha < 1$, then the service-time distribution has infinite mean, which implies that the waiting-time process fails to have a proper steady-state distribution, in particular,

$$W_k \to \infty \quad \text{as} \quad k \to \infty \quad \text{w.p.1.}$$

9.7.1. Heavy-Traffic Limits

We can use the heavy-traffic stochastic-process limits to show how the waiting times should grow over finite time intervals. Similar limits will hold for the queue-length processes. For these limits, it suffices to consider a single queueing system. Since the mean service time is infinite, the traffic intensity is infinite here, and thus plays no role. Let the random elements of D be defined by

$$\begin{aligned} \mathbf{S}_n(t) &\equiv n^{-1/\alpha} S_{\lfloor nt \rfloor}, \\ \mathbf{W}_n(t) &\equiv n^{-1/\alpha} W_{\lfloor nt \rfloor}, \quad t \geq 0 \ . \end{aligned} \tag{7.2}$$

Theorem 9.7.1. (service times with very heavy tails) *Consider the standard single-server queue with interarrival times and service times coming from a sequence of IID random vectors* $\{(U_k, V_k)\}$. *Suppose that* $EU_1 < \infty$ *and* (7.1) *holds with* $0 < \alpha < 1$. *Then*

$$(\mathbf{S}_n, \mathbf{W}_n) \Rightarrow (\mathbf{S}, \mathbf{S})) \quad \textit{in} \quad D([0,\infty), \mathbb{R}^2, SJ_1) \ ,$$

for $\mathbf{S}_n$ *and* $\mathbf{W}_n$ *in* (7.2), *where* $\mathbf{S}$ *is the* α*-stable Lévy motion with* $\beta = 1$ *characterized by Theorems* 4.5.2 *and* 4.5.3 *that arises as the stochastic-process limit for partial sums of the service times alone.*

Proof. Since $EU_1 < \infty$, $\{U_k : k \geq 1\}$ obeys the strong law of large numbers, which in turn implies a functional strong law; see Corollary 3.2.1 in the Internet Supplement. Hence the FCLT for $\mathbf{S}_n$ follows for the FCLT for the partial sums of the service times alone (without translation term), by virtue of Theorem 11.4.5. The FCLT for the service times alone follows from Theorems 4.5.2 and 4.5.3. We obtain the limit theorem for the scaled waiting times by applying the continuous-mapping approach with the reflection map; in particular, we can apply Theorem 9.3.1 (a). Finally, the limit process $\mathbf{W}$ has the indicated form because $\mathbf{S}$ has nondecreasing sample paths since $\beta = 1$. ∎

As a consequence, of Theorem 9.7.1, we can approximate the transient waiting times by

$$W_k \approx n^{1/\alpha} \mathbf{S}(k/n) \ , \quad k \geq 0 \ ,$$

for any k.

We now show that we can calculate the pdf and cdf of the $S_\alpha(\sigma, 1, 0)$ stable distribution for $0 < \alpha < 1$. Paralleling the case $\alpha = 3/2$ described in Theorem 8.5.4, the case $\alpha = 1/2$ is especially tractable. As noted in

Section 4.5, for $\alpha = 1/2$, we obtain the Lévy distribution; i.e., the $S_\alpha(1,1,0)$ distribution has cdf

$$G_{1/2}(x) = 2\Phi^c(1/\sqrt{x}), \quad x \geq 0$$

where $\Phi^c(x) \equiv P(N(0,1) > x)$ and pdf

$$g_{1/2}(x) = \frac{1}{\sqrt{2\pi x^3}} e^{-1/2x}, \quad x \geq 0 \ ;$$

see p. 52 of Feller (1971).

More generally, we can apply numerical inversion of Laplace transforms again to calculate the pdf and ccdf of the stable subordinator $\mathbf{S}(t)$. We exploit the fact that the distribution $S_\alpha(\sigma,1,0)$ of $S^\alpha(1)$ has support on the positive halfline. That makes the bilateral Laplace transform in (5.17) in Section 4.5 a bonafide Laplace transform. We exploit self-similarity to relate the distribution at any time t to the distribution at time 1, i.e.,

$$\mathbf{S}(t) \stackrel{\mathrm{d}}{=} t^{1/\alpha}\mathbf{S}(1) \ . \tag{7.3}$$

Hence it suffices to consider the single-parameter family of distributions $S_\alpha(1,1,0)$.

By (5.12) in Section 4.5, we know that the ccdf of $S_\alpha(1,1,0)$ decays as $x^{-\alpha}$. Hence, for $0 < \alpha < 1$, the random variable $\mathbf{S}(t)$ has infinite mean. By (7.3), we expect $\mathbf{S}(t)$ to grow like $t^{1/\alpha}$ as t increases. However, we should expect much of the growth to be in large jumps. To illustrate the form of the ccdf's, we give the ccdf values of $S_\alpha(1,1,0)$ for three values of α in Table 9.1, again computed by numerical transform inversion, exploiting the Euler algorithm in Abate and Whitt (1995a).

The cdf of the stable law $S_\alpha(1,1,0)$ reveals the consequences of the heavy tail, but it does not directly show the jumps in the stable Lévy motion. We see the jumps more directly when we consider the first passage times to high levels. We can exploit the convergence to a stable Lévy motion to show, asymptotically, how the waiting-time process reaches new levels when the service-time distribution has such a heavy tail (with $0 < \alpha < 1$).

9.7.2. First Passage to High Levels

As observed in Section 4.5, a stable Lévy motion with $0 < \alpha < 2$ is a pure-jump stochastic process. Thus, the stable Lévy motion passes any specified level by making a jump. (See Bertoin (1996).) Hence the process is below the level just before the jump and above the level immediately

x	G_α^c ccdf of $S_\alpha(1,1,0)$		
	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
$(0.01)2^0 = 0.01$	0.9037	1.0000	1.0000
$(0.01)2^1 = 0.02$	0.8672	1.0000	1.0000
$(0.01)2^2 = 0.04$	0.8251	0.9996	1.0000
$(0.01)2^4 = 0.16$	0.7282	0.9229	1.0000
$(0.01)2^6 = 0.64$	0.6233	0.6232	0.7371
$(0.01)2^8 = 2.56$	0.5197	0.3415	0.1402
$(0.01)2^{10} = 10.24$	0.4242	0.1749	0.3739 $e-1$
$(0.01)2^{12} = 40.96$	0.3404	0.8798 $e-1$	0.1154 $e-1$
$(0.01)2^{16} = 655.36$	0.2112	0.2204 $e-1$	0.1220 $e-2$
$(0.01)2^{20} = 10486$	0.1269	0.5510 $e-2$	0.1324 $e-3$
$(0.01)2^{24} = 167,772$	0.7477 $e-1$	0.1377 $e-2$	0.1440 $e-4$
$(0.01)2^{28} = 2,684,000$	0.4359 $e-1$	0.3444 $e-3$	0.1567 $e-5$
$(0.01)2^{32} = 42,949,000$	0.2525 $e-1$	0.8609 $e-4$	0.1705 $e-6$

Table 9.1: Tail probabilities of the stable law $S_\alpha(1,1,0)$ for $\alpha = 0.2$, 0.5 and 0.8 computed by numerical transform inversion.

after the jump. It is significant that we can obtain useful characterizations of the distributions of the values immediately before and after first passing any level for the limiting stable Lévy motion. We describe the asymptotic distribution of the last value before the jump as the level increases.

Stochastic-process limits for these quantities follow from the continuous-mapping approach with Theorem 13.6.5. Explicit expressions for the distributions associated with the limiting stable Lévy notion follow from the generalized arc sine laws; see Sections III and VIII of Bertoin (1996).

For $z > 0$, let τ_z be the *first passage time* to a level beyond z; i.e., for $x \in D$,

$$\tau_z(x) \equiv x^{-1}(z) \equiv \inf\{t \geq 0 : x(t) > z\} \tag{7.4}$$

with $\tau_z(x) = \infty$ if $x(t) \leq z$ for all t. Let γ_z be the associated *overshoot*; i.e.,

$$\gamma_z(x) = x(\tau_z(x)) - z \ . \tag{7.5}$$

Let λ_z be the *last value* before the jump; i.e.,

$$\lambda_z(x) \equiv x(\tau_z(x)-) \ . \tag{7.6}$$

Let these functions also be defined for the discrete-time process $W \equiv \{W_k\}$ (without scaling) in the same way.

Note that the scale parameter σ enters in simply to the first passage time, i.e., for $y > 0$

$$\tau_z(\mathbf{S}(y\cdot)) = y^{-1}\tau_z(\mathbf{S}) \ ,$$

and does not appear at all in the overshoot or the last value before passage (because σ corresponds to a simple time scaling).

Also note that we can determine the distribution of the overshoot and the jump size for the waiting times in a GI/GI/1 model if we know the distribution of the last value $\lambda_z(W)$: Because of the IID assumption for $\{(U_k, V_k)\}$,

$$P(\gamma_z(W) > x | \lambda_z(W) = y) = P(V_1 - U_1 > x + z - y | V_1 - U_1 > z - y) \ . \quad (7.7)$$

When x and $z-y$ are both large, we can exploit the service-time tail asymptotics in (7.1) to obtain the useful approximation

$$\begin{aligned} &P(V_1 - U_1 > x + z - y | V_1 - U_1 > z - y) \\ &\quad \approx P(V_1 > x + z - y | V_1 > z - y) \\ &\quad \approx ((z-y)/(x+z-y))^\alpha \ . \end{aligned} \quad (7.8)$$

hence, much interest centers on determining the distribution of the last value before the jump for the waiting times. That exact distribution is hard to come by directly, so that the heavy-traffic limit is helpful.

Theorem 9.7.2. (limits for the first-passage time, overshoot and last value) *Under the conditions of Theorem* 9.7.1, *for* $z > 0$,

$$n^{-1}\tau_{zn^{1/\alpha}}(W) \Rightarrow \tau_z(\mathbf{S}) \quad in \quad \mathbb{R} \quad as \quad n \to \infty \ ,$$

so that

$$\lim_{n\to\infty} P(\tau_{zn^{1/\alpha}}(W) > nx) = P(\mathbf{S}(x) \le z) \ ;$$

$$n^{-1/\alpha}\gamma_{zn^{1/\alpha}}(W) \Rightarrow \gamma_z(\mathbf{S}) \quad in \quad \mathbb{R} \quad as \quad n \to \infty \ ,$$

so that, for $b > z$,

$$\begin{aligned} &\lim_{n\to\infty} P(\gamma_{zn^{1/\alpha}}(W) > (b-z)n^{1/\alpha}) \\ &\qquad = \frac{1}{\Gamma(\alpha)\Gamma(1-\alpha)} \int_0^z x^{\alpha-1}(b-x)^{-\alpha} dx \ ; \end{aligned} \quad (7.9)$$

$$n^{-1/\alpha} W(\lfloor \tau_{zn^{1/\alpha}}(W) - \rfloor) \Rightarrow \mathbf{S}(\tau_z(\mathbf{S})-)$$

and, for $0 < b < 1$,

$$\lim_{z\to\infty} P(\mathbf{S}(\tau_z(\mathbf{S})-) > zb) = \int_b^1 \frac{\sin(\alpha\pi)dt}{\pi t^{1-\alpha}(1-t)^\alpha} \ . \tag{7.10}$$

Proof. By Theorem 13.6.5, the first-passage time, overshoot and last-value functions are almost surely continuous functions on D with respect to the limiting stable Lévy motion. Hence we can apply the continuous mapping theorem, Theorem 3.4.3. Note that

$$n^{-1}\tau_{zn^{1/\alpha}}(W) = \tau_z(n^{-1/\alpha}W_{\lfloor n\cdot\rfloor}) \ ,$$

$$n^{-1/\alpha}\gamma_{zn^{1/\alpha}}(W) = n^{-1/\alpha}W_{\lfloor \tau_{zn^{1/\alpha}}(W)\rfloor} - z = \gamma_z(n^{-1/\alpha}W_{\lfloor n\cdot\rfloor})$$

and

$$n^{-1/\alpha} W(\lfloor \tau_{zn^{1/\alpha}}(W) - \rfloor) = n^{-1/\alpha} W_{\lfloor n\tau_z(n^{-1/\alpha}W(n\cdot)-)\rfloor} \ .$$

For (7.9), see Exercise 3, p. 238, and p. 241 of Bertoin (1996). For (7.10), see Theorem 6, p. 81, of Bertoin (1996). ∎

The limiting distribution in (7.10) is called the generalized arc sine law. Its density is in general an asymmetric U-shaped function. The case $\alpha = 1/2$ produces the standard arc sine density in Corollary 4.3.1. Consistent with intuition, as α decreases, the chance of the scaled last value being relatively small (making the final jump large for a large level z) increases.

9.8. An Increasing Number of Arrival Processes

In this section we establish heavy-traffic limits for queues with superposition arrival processes, where the number of arrival processes being superposed increases in the limit. Related results for fluid queues were established in Section 8.7.

9.8.1. Iterated and Double Limits

From Theorem 9.4.1, we see that the FCLT for a superposition of m IID counting processes is the same as the FCLT for a single counting process, except for the obvious deterministic scaling. Indeed, in Section 9.6 we observed that the dimensionless variability parameter c_{HT}^2 defined in (6.9) and (6.10) is independent of m.

However, we obtain a different picture from the fundamental limit for superpositions of point processes, where the number of superposed processes gets large with the total rate held fixed. (That limit is sometimes called the law of small numbers.) Then the superposition process converges to a Poisson process; e.g., see Çinlar (1972) or Daley and Vere Jones (1988). For this limit, convergence in (D, J_1) is equivalent to convergence of the finite-dimensional distributions.

Theorem 9.8.1. (Poisson limit for superposition processes) *Suppose that A^i are IID counting processes with stationary increments and without multiple points (all jumps in A^i are of size 1). Then*

$$A_m \Rightarrow A \quad in \quad (D, J_1) \quad as \quad m \to \infty \ , \tag{8.1}$$

where

$$A_m(t) \equiv \sum_{i=1}^{m} A^i(t/m), \quad t \geq 0, \quad m \geq 1 \, , \tag{8.2}$$

and A is a homogeneous Poisson process with intensity

$$\lambda \equiv E[A(t+1) - A(t)] = E[A^1(t+1) - A^1(t)] \ . \tag{8.3}$$

In view of Theorem 9.8.1, we might well expect the superposition arrival process for large m to behave like a Poisson process in the heavy-traffic limit. However, if A^1 is a Poisson process, then $\mathbf{S}^u_{n,1} \Rightarrow \mathbf{S}^u_1$ with $c_n = \sqrt{n}$ and $\mathbf{S}^u_1$ a standard Brownian motion; i.e., the dimensionless variability parameter is $c^2_U = 1$. Clearly, Theorem 9.4.1 does not capture this Poisson tendency associated with large m. The two iterated limits $\lim_{\rho \to 1} \lim_{m \to \infty}$ and $\lim_{m \to \infty} \lim_{\rho \to 1}$ are not equal. The reason that these iterated limits do not coincide is that the superposition process looks different in different time scales. The iterated limits do not agree because the Poisson superposition limit focuses on the short-time behavior, while the heavy-traffic limit focuses on long-time behavior.

Remark 9.8.1. *Different variability at different time scales.* A Poisson process is relatively simple in that it tends to have the same level of variability at all time scales. For example, if A is a Poisson counting process with rate λ, then both the mean and the variance of $A(t)$ are λt for all $t > 0$. In contrast, a superposition of a large number m of IID stationary point processes (without multiple points), where each component process is not nearly Poisson, tends to have different levels of variability at different

time scales. Consistent with Theorem 9.8.1, for large m, the superposition process tends to look like a Poisson process in a short time scale, but it looks like a single component point process in a long time scale.

For example, for a superposition of IID point processes with large m and small t, the variance of $A(t)$ tends to be approximately λmt, where λ is the rate of one component process, just as if A were a Poisson process. However, consistent with Theorem 9.4.1, under regularity conditions, for any given m, the variance of $A(t)$ approaches $\lambda c_a^2 mt$, where

$$c_a^2 = \lim_{t\to\infty} Var(A_1(t)/\lambda t) ,$$

with A_1 being the counting process of one source. If A_1 is a Poisson process, then $c_a^2 = 1$, but more generally c_a^2 can be very different from 1.

The heavy-traffic limits in Section 9.4 for queues with a superposition of a fixed number of component processes capture only the large-time-scale variability of the superposition process. That is appropriate for the queue for any number m of component processes provided that the traffic intensity ρ is large enough. However, in practice ρ may not be large enough.

The problem, then, is to understand how variability in the input, with levels varying over different time scales, affects queueing performance. Consistent with intuition, it can be shown that the large-time-scale variability tends to determine queue performance at very high traffic intensities, while the short-time-scale variability tends to determine queue performance at very low traffic intensities. More generally, we conclude that variability at longer times scales become more important for queue performance as the traffic intensity increases. See Section 9.9, Sriram and Whitt (1986), Fendick, Saksena and Whitt (1989, 1991) and Fendick and Whitt (1989) for more discussion. As shown by Whitt (1985a), for superposition processes, we gain insight into the effect of different variability at different time scales upon queueing performance by considering the double limit as $\rho \uparrow 1$ and $m \to \infty$. ∎

In order to have a limit that captures some of the structure of the superposition process not seen in either a single component process or the Poisson process, we consider a double limit, letting the number of component processes be n, and then letting $\rho_n \uparrow 1$ as $n \to \infty$. As in Theorem 9.8.1, we rescale time in the superposition process so that the total arrival rate is fixed, say at 1. Thus the superposition arrival process alone approaches a rate-1 Poisson process as the number n of components increases. In the n^{th} queueing model with n component arrival processes, we let the service times

have mean ρ_n^{-1}, so that the traffic intensity in model n is ρ_n. We achieve heavy traffic by letting $\rho_n \uparrow 1$ as $n \to \infty$.

The double limit considered in this section has advantages and disadvantages. Its first advantage is that it may more faithfully describe queues with superpositions of a large number of component arrival processes. Its second advantage is that, even if the interarrival times have heavy-tailed distributions, the limit process is likely to have continuous sample paths. However, a disadvantage is that the limit process is more complicated, because it does not have independent increments.

We start by considering the superposition arrival process alone. Treating the arrival process alone, we first scale time by n^{-1} to keep the rate fixed. Then we scale time again by n to establish the FCLT. These two time scalings cancel out, so that there is no time scaling inside the arrival process. In particular, the scaled arrival process is

$$\begin{aligned} \mathbf{A}_n(t) &\equiv c_n^{-1}(A_n(nt/n) - \lambda nt) \\ &= c_n^{-1}(A_n(t) - \lambda nt) \\ &= c_n^{-1}\left(\sum_{i=1}^{n} A^i(t) - \lambda nt\right) \\ &= c_n^{-1}\sum_{i=1}^{n}(A^i(t) - \lambda t), \quad t \geq 0\,. \end{aligned} \tag{8.4}$$

From the final line of (8.4), we see that the final scaled process $\mathbf{A}_n$ can be represented as the scaled sum of the IID processes $\{A^i(t) - \lambda t : t \geq 0\}$. Thus limits for $\mathbf{A}_n$ follow from the CLT for processes in Section 7.2.

Now, following and extending Whitt (1985a), we establish a general heavy-traffic stochastic-process limit for a queue with a superposition arrival process, where the number of component arrival processes increases in the limit. (See Knessl and Morrison (1991), Kushner and Martins (1993, 1994), Kushner, Yang and Jarvis (1995), Brichet et al. (1996, 2000) and Kushner (2001) for related limits.) We consider the general space scaling by c_n, where $c_n \to \infty$ and $n/c_n \to \infty$.

Theorem 9.8.2. (general heavy-traffic limit for a queue with a superposition arrival process having an increasing number of components) *Consider a sequence of single-server queueing models indexed by n, where the service times are independent of the arrival times and the arrivals come from the superposition of n IID component arrival processes A^i. Suppose that*

$$\mathbf{S}_n^v \Rightarrow \mathbf{S}^v \quad \textit{in} \quad (D, M_1) \tag{8.5}$$

for $\mathbf{S}_n^v$ *in* (3.1), $P(\mathbf{S}^v(0) = 0) = 1$, *and*

$$P(t \in Disc(\mathbf{S}^v)) = 0 \quad \textit{for all} \quad t \ . \tag{8.6}$$

Suppose that

$$\mathbf{A}_n \Rightarrow \mathbf{A} \quad \textit{in} \quad (D, M_1) \tag{8.7}$$

for $\mathbf{A}_n$ *in* (8.4), $P(\mathbf{A}(0) = 0) = 1$ *and*

$$P(t \in Disc(\mathbf{A})) = 0 \quad \textit{for all} \quad t \ . \tag{8.8}$$

If $c_n \to \infty$, $n/c_n \to \infty$ *and*

$$\eta_n \equiv n(\mu_n^{-1} - \lambda^{-1})/c_n \to \eta \quad \textit{as} \quad n \to \infty \ , \tag{8.9}$$

then the conditions and conclusions of Theorems 9.3.3 *and* 9.3.4 *hold with* $\lambda_n = \lambda$, $\mathbf{S}^u = \lambda^{-1}\mathbf{A} \circ \lambda^{-1}\mathbf{e}$ *and the* WM_1 *topology on the product space* D^k.

Proof. It is easy to verify that the conditions here imply the conditions in Theorems 9.3.3 and 9.3.4: First, we can apply Theorem 7.3.2 to get

$$\mathbf{S}_n^u \Rightarrow \mathbf{S}^u = -\lambda^{-1}\mathbf{A} \circ \lambda^{-1}\mathbf{e} \ .$$

Then we can apply Theorem 11.4.4 to get

$$(\mathbf{S}_n^u, \mathbf{S}_n^v) \Rightarrow (\mathbf{S}^u, \mathbf{S}^v) \quad \text{in} \quad (D^2, WM1) \ .$$

Conditions (8.6) and (8.8) imply condition (3.10). The conditions also imply (3.19). ∎

Remark 9.8.2. *The case of a Lévy counting process.* If A is a Lévy process, then the scaled superposition process in (8.4) satisfies

$$\mathbf{A}_n(t) \stackrel{\mathrm{d}}{=} c_n^{-1}[A^1(nt) - \lambda nt], \quad t \geq 0 \ ,$$

as in (3.4) with constant λ, using the reasoning in Remark 9.4.1. In that case, Theorem 9.8.2 adds nothing new. The counting process A is a Lévy process if it is a Poisson process or, more generally, a batch Poisson process, with the batches coming from a sequence of IID integer-valued random variables. The scaled batch-Poisson process can converge to a non-Brownian stable Lévy motion. ∎

We now focus on the way the number of sources, n, and the traffic intensity, ρ, should change as $n \to \infty$ and $\rho \uparrow 1$. For that purpose, suppose that $c_n = n^H$ for $0 < H \leq 1$, then (8.9) implies that

$$n^{1-H}(1 - \rho_n) \to |\lambda\eta| \quad \text{as} \quad n \to \infty \ . \tag{8.10}$$

As the component arrival processes get more bursty, H increases. As H increases, n^{1-H} increases more slowly as a function of n. Thus, with greater burstiness, ρ can approach 1 more slowly to have the heavy-traffic limit in Theorem 9.8.2.

We now have criteria to determine when the two iterated limits tell the correct story: If

$$n >> (|\lambda\eta|/(1-\rho))^{1/(1-H)} \ ,$$

then the arrival process should behave like a Poisson process in the heavy-traffic limit; if

$$n << (|\lambda\eta|/(1-\rho))^{1/(1-H)} \ ,$$

then the arrival process should behave like a single component arrival process in the heavy-traffic limit. The intermediate case covered by (8.10) is more complicated.

In Section 7.2 we have given sufficient conditions for the condition $\mathbf{A}_n \Rightarrow \mathbf{A}$ in (8.7). We illustrate by giving a result from Whitt (1985a) for superpositions of renewal processes, drawing on Theorem 7.2.3.

Theorem 9.8.3. (reflected Gaussian heavy-traffic limit for a queue with a superposition arrival process having an increasing number of renewal components) *Consider a sequence of single-server queueing models indexed by n, where the service times are independent of the arrival times and the arrivals come from the superposition of n IID component stationary renewal processes A^i with interrenewal cdf F having mean λ^{-1}. Suppose that*

$$\mathbf{S}_n^v \Rightarrow \mathbf{S}^v \quad \textit{in} \quad (D, J_1) \tag{8.11}$$

for $\mathbf{S}_n^v$ in (3.1) with $c_n = \sqrt{n}$ and $\mathbf{S}^v$ a zero-mean Brownian motion. Suppose that

$$\lim_{t \to 0} t^{-1}(F(t) - F(0)) < \infty \ . \tag{8.12}$$

Suppose that

$$\eta_n = \sqrt{n}(\mu_n^{-1} - \lambda^{-1}) \to \eta \quad \textit{as} \quad n \to \infty \ . \tag{8.13}$$

Then

$$\mathbf{A}_n \Rightarrow \mathbf{A} \quad \textit{in} \quad (D, J_1) \tag{8.14}$$

where $\mathbf{A}$ *is a zero-mean Gaussian process with stationary increments and continuous sample paths. The limit process* $\mathbf{A}$ *has the covariance function of* A^1*, which is characterized in Theorem* 7.2.4*. Then the conditions and conclusions of Theorems* 9.3.3 *and* 9.3.4 *hold with* $c_n = \sqrt{n}$*,* $\lambda_n = \lambda$ *and*

$$\mathbf{S}^u = -\lambda^{-1}\mathbf{A} \circ \lambda^{-1}\mathbf{e} \ . \tag{8.15}$$

Consequently, the limit processes $\mathbf{S}$ *and* $\mathbf{X}$ *are Gaussian processes with stationary increments and continuous sample paths.*

Proof. Apply Theorems 9.3.3, 9.3.4 and 7.2.3. ∎

Unfortunately the limit processes for the waiting time, queue-length and workload processes stemming from Theorem 9.8.3 are relatively intractable, because the limit processes $\mathbf{S}$ and $\mathbf{X}$ here do not have independent increments. However, since $\mathbf{S}$ and $\mathbf{X}$ are Gaussian processes, we can obtain approximations for the steady-state distributions of the queueing-content limit processes $\mathbf{W}$, $\mathbf{L}$, $\mathbf{Q}$ and $\mathbf{Q}^A$ by applying Section 8.8. We can also establish a second limit to RFBM as in Section 8.7.

9.8.2. Separation of Time Scales

When we let the number of sources become large in a single-server queue, we change the relevant time scales of the sources relative to the server. With n IID sources, the interarrival times for each source become of order $O(n)$, while the service times remain of order $O(1)$. When we scale time by n for the heavy-traffic limit, the interarrival times for each source become of order $O(1)$, while the service times become of order $O(n^{-1})$. From either perspective, the interarrival times for each source are of order $O(n)$ times longer than the service times. Thus, as n increases, the relevant time scales for the sources and the server separate. Consequently, for large n, the small-time-scale behavior of the source arrival processes (from their own perspective) can significantly affect the large-time-scale or heavy-traffic behavior of the queue.

Consistent with that observation, the limit process $\mathbf{A}$ in Theorem 9.8.3 providing the contribution of the arrival process to the heavy-traffic limit depends on the component process A_1 through its correlation function. Thus, unlike the case of a single source, locally smoothing the input for each source with many sources can dramatically reduce the congestion in heavy traffic. In contrast, for a single source, the heavy-traffic behavior of the queue depends on the source arrival process only through its CLT behavior, which of course depends on the large-time-scale behavior of that source.

As noted by Whitt (1988), there is a separation of time scales for flows in multi-class queueing networks: When one source at a queue has an arrival rate much smaller than the service rate (usually because the server is shared by many sources), the departure process for that class tends to be very similar to the arrival process for that class, because the service and delay experienced at that queue tend to be in a shorter time scale. Thus, in a flow through a network from source to destination, where the arrival rate of that flow is much smaller than the service rate at all queues on its path, the arrival process at the destination will be very similar to the original flow emitted from the source. This property has been further exposed by Wischik (2001b) using moderate-deviation limits.

For discussions about time scales in queues associated with communication networks, see Sriram and Whitt (1986), Fendick and Whitt (1989), Tse, Gallager and Tsitsiklis (1995), Jelenković, Lazar and Semret (1997), Grossglauser and Tse (1999), Greenberg, Srikant and Whitt (1999) and Srikant and Whitt (2001). ∎

Example 9.8.1. *Token-bank rate-control throttles.* The separation of time scales has implication for the effectiveness of devices to regulate traffic. One such device is a token-bank rate-control throttle; see Berger (1991), Berger and Whitt (1992a, b, 1994, 1995b) and references cited there. The operation of such a throttle is depicted in Figure 9.8.1. The throttle contains two finite buffers, one for jobs and one for tokens. The jobs may be packets in a high-speed packet network or call-setup requests in a telecommunications switching system. The buffer for tokens, called a token bank, is typically a fictitious buffer, because the token bank is usually implemented by a counter with a cap, but it is convenient to think of physical tokens. These tokens arrive deterministically and evenly spaced from an infinite source.

Tokens that arrive to a full token bank are blocked and lost. If the bank contains a token when a job arrives to the throttle, then the job is allowed to pass through, and one token is removed from the token bank. If the token bank does not contain any tokens when a job arrives, then the job queues in the job buffer if the job buffer is not full. If a job arrives to find a full job buffer, then the job is not admitted and is said to have "overflowed". In packet networks, the overflowed packet may be discarded or may be marked and later treated as a lower priority class.

The token-bank rate-control throttle is closely related to the *leaky-bucket regulator*. With the conventional definition, the leaky bucket has a constant *drain rate* r and a capacity C. At a job arrival, if the bucket content is below $C-1$, then the job is admitted and the bucket content is increased by

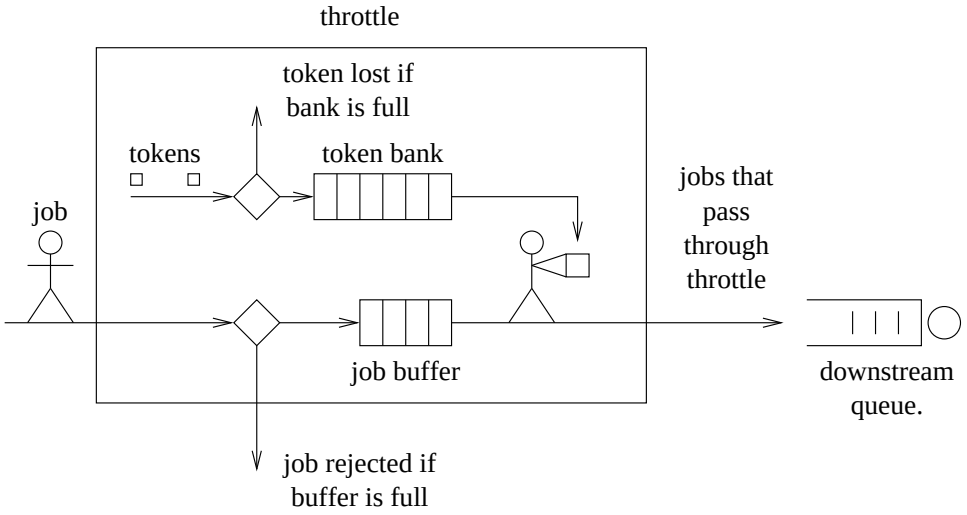

Figure 9.1: Diagram of a token-bank rate-control throttle with a job-buffer regulating traffic to a downstream queue.

1. Otherwise, the job overflows. The bucket drains out at a deterministic rate r. When the bucket is empty, the draining stops. The draining process starts again upon the next job arrival. The arrival brings the bucket content to 1, and a new busy period of the bucket begins. Thus, the time epochs at which a unit of content drains out of the bucket do not remain synchronous in time, but instead experience a phase shift each time the bucket empties.

In contrast, with a token-bank rate-control throttle, the token arrival process continues to run independent of the state of the bank, so that the token arrival epochs do remain synchronous for all time. The leaky bucket is equivalent to a modified rate-control throttle, without job buffer, in which the deterministic token arrival process stops whenever the token bank becomes full, and starts again at the next job arrival epoch. Just like the rate-control throttle, the leaky bucket can be supplemented by adding a job buffer. Hence, our remarks here about token-bank rate-control throttles also apply to leaky-bucket regulators.

An important initial observation for understanding the performance of the throttle is the *overflow invariance property* established by Berger (1991) and Berger and Whitt (1992a): Except for a finite initial period to count for initial conditions, the job overflow process depends on the (finite) capacity of the token bank, C_T, and the (finite) capacity of the job buffer, C_J, only via

their sum $C = C_T + C_J$. The overflow invariance property implies that we can decompose the question about the performance of the throttle into two separate parts: First, there is the traffic shaping caused by job rejections, which depends only on the total capacity C. Second, there is the additional traffic shaping provided by a job buffer given a fixed total capacity C.

The traffic shaping caused by job rejections can be studied by establishing heavy-traffic limits for the throttle; that was done by Berger and Whitt (1992b). Following Berger and Whitt (1992a, 1994), here we will focus on the second question: What is the traffic-shaping benefit provided by the job buffer, given fixed total capacity C?

For given total capacity C, we should prefer no job buffer ($C_J = 0$) if there were no performance differences, because a job buffer is an actual buffer requiring resources to implement. The reason for having a job buffer is that it can provide additional traffic shaping. The potential advantages of a job buffer are easy to see when we contrast an all-token-bank throttle (with $C_J = 0$) to an all-job-buffer throttle (with $C_T = 0$). With an all-token-bank throttle, the throttle can admit batches of jobs of size C. In contrast, with an all-job-buffer throttle, the successive admission epochs of jobs are always separated by at least the deterministic interval between successive token arrivals.

Early proponents of rate-control throttles with job buffers noted the smoothing properties of the throttle. For example, they showed that the throttle reduced the variability (e.g., as measured by the squared coefficient of variation) of the stationary interval between successive job admission epochs. However, through stochastic analysis and simulation, Berger and Whitt (1992a, 1994) showed that, while the traffic smoothing benefit of the throttle was dramatic in a short time scale, it was much less so in a long time scale. Indeed, they showed that the heavy-traffic limiting behavior at a downstream queue fed by a source with a rate-control throttle is independent of the job buffer, given fixed total capacity. More generally, simulations showed that the job buffer tends to provide only a relatively minor reduction of congestion in a downstream queue.

However, most systems actually have traffic from many sources entering the downstream system. As noted above, when the number of sources increases, the short-time behavior of the individual sources begins to have impact upon the large-time-scale behavior of the queue. In the limit, there is a separation of time scales. Consistent with that observation, simulations show that, in marked contrast to the case of a queue fed by a single source, the job buffer provides a dramatic smoothing benefit when 100 identical sources regulated by throttles feed a downstream queue. The separation of

time scales provided by many sources makes the short-time-scale behavior of the individual sources relevant to the large-time-scale behavior of the queue.

Consistent with that observation, the simulations also show that the synchronization of many token arrival streams can be a major source of congestion: If there are many sources, and the token arrival epochs of these sources are synchronous, then there can be bursts of arrivals at each token arrival epoch. This effect tends not to appear, however, if all the throttles are not synchronized, i.e., if the phase is random for each throttle.

For recent work focusing on the impact of rate control throttles on long-range dependent input, see Vamvakos and Anantharam (1998) and Gonzáles-Arévalo and Samorodnitsky (2001). ∎

9.9. Approximations for Queueing Networks

Most systems experiencing congestion are not single queues, but networks of queues. However, a cardinal principle of performance analysis is: *Look for the bottleneck!* Often there is a critical resource that primarily determines system performance. When viewed correctly, the complex queueing network often reduces to a smaller system that is easier to analyze. Indeed, it often suffices to consider a single queue.

Hence, from a practical perspective, there is much justification for emphasizing single queues. However, it is also helpful to be able to analyze queueing networks.

9.9.1. Parametric-Decomposition Approximations

In this section we discuss heuristic approximations for queueing networks. These approximation are called *parametric-decomposition approximations* because the queues are analyzed separately after approximately characterizing the arrival process at each queue by two parameters; see Whitt (1983a,b, 1995) and Buzacott and Shanthikumar (1993). (The first work in this direction was done by Reiser and Kobayashi (1974), Sevcik et al. (1977) and Kuehn (1979).)

We discuss parametric-decomposition approximations here because heavy-traffic limits can play an important role in choosing appropriate variability parameters. Indeed, important insight is provided by the heavy-traffic limit for a queue with a superposition arrival process, where the number of component arrival processes increases in the heavy traffic limit, just considered in the previous section.

With parametric-decomposition approximations, the goal is to obtain improved performance predictions compared to more elementary one-parameter models such as the $M/M/1$ queue and the single-class open Jackson queueing network (a network of $M/M/1$ queues with Markovian routing); see Jackson (1957, 1963). Variability has an impact on the performance of these one-parameter models, but they provide no parameters to quantify the degree of variability.

In this section we are primarily interested in exploiting heavy-traffic limits to improve the quality of parametric-decomposition approximations. Along the way, we point out significant difficulties, where initial simple approaches break down. When considering approximation errors, it is good to keep in mind that in engineering applications the error in model fit is usually larger than the error in approximating the solution of the model.

We start by considering how to approximately characterize the distribution of a nonnegative real-valued random variable. It is natural to partially characterize the distribution by its mean and squared coefficient of variation (SCV, variance divided by the square of the mean). Thus it is natural to partially characterize a renewal process by the mean and SCV of the interrenewal time.

However, it is difficult to adequately characterize a general stationary arrival process by only two parameters, because in addition to the general interarrival-time distribution, there may be complicated dependence among the interarrival times. For models, the arrival rate can be determined exactly. The difficulty is in finding an appropriate second parameter to characterize the variability.

The variability of a general stationary arrival process often looks different in different time scales. Consequently, the variability impact on the congestion in a following queue often depends on the traffic intensity of that queue. Hence, following Whitt (1995), in order to partially characterize a general stationary arrival process, we propose using the arrival rate and a *variability function* that gives a variability parameter as a function of the traffic intensity in a following queue. When that arrival process appears in a queue, we obtain a variability parameter by evaluating the variability function at the traffic intensity of the queue. (Fendick and Whitt (1989) investigate how variability as a function of *time* in an arrival process can be converted into variability as a function of the *traffic intensity* in a following queue.)

In a typical queueing-network application, there are multiple classes of customers, each with their own arrival, service and routing pattern. It is often realistic to assume that the routing is primarily deterministic for each

class, so we will consider the case of deterministic routing. With deterministic routing, there is an exogenous arrival process to some queue for each class, which we will regard as a renewal process partially characterized by the mean and the SCV of an interrenewal time. (We will later discuss extensions to non-renewal arrival processes partially characterized by variability functions.) Each customer visits a sequence of queues in the network, possibly returning to the same queue more than once, and then leaves the network. At each queue on the customer's route, there is a service-time distribution, which is partially characterized by its mean and SCV. It is assumed that the arrival process and the service times are mutually independent. The service-time distributions may differ at different queues. The service-time distributions also may differ at the same queue for different customers or even for the same customer upon different visits to that queue.

The model data for one customer class might be the vector

$$(1, 2, 4; 2, 1, 0; 3, 1, 1; 2, 5, 1) \ . \tag{9.1}$$

The first triple describes the exogenous arrival process: Customers from that class enter the network at queue 1 with an exogenous renewal arrival process having arrival rate 2 and interarrival-time SCV 4. Afterwards, these customers visit queues 2, 3 and 2, in that order, and then leave the network. On the first visit to queue 2, the service time has mean 1 and SCV 0, while on the second visit to queue 2 the service time has mean 5 and SCV 1.

We must also specify the queues. For simplicity, we will consider only single-server queues with unlimited waiting room and the FCFS service discipline, but clearly the general approach can accomodate a wide variety of queues.

Given the specified model data partially characterizing the arrival and service processes of each customer class in an open queueing network of single-server FCFS queues, the goal is to describe the performance. We want to determine approximate queue-length distributions at each queue and approximate sojourn-time (time-in-system) distributions for each customer class.

Here we will only discuss the mean (steady-state) sojourn time for one class. The mean sojourn time is the sum of the mean waiting times (before beginning service) and the mean service times at all the queues on the customer's route. The mean service times are directly specified for each class as part of the model data, so we use them. (To do otherwise could introduce large errors unnecessarily.) In general, the waiting-time distribution and its mean can depend on the customer class, but we will use an approximation

for the mean waiting time for an arbitrary customer at that queue. Hence, here our goal reduces to developing an approximation for the mean waiting time for an arbitrary customer at each queue in the network.

To determine the approximate mean waiting time at any single queue, we act as if we have a $GI/GI/1$ queue partially characterized by the mean λ^{-1} and SCV c_a^2 of an interarrival time and the mean μ^{-1} and SCV c_s^2 of a service time. We will use the heavy-traffic approximation, refined by the exact $M/GI/1$ formula, namely,

$$EW \equiv EW(\lambda, c_a^2, \mu, c_s^2) \approx \frac{\mu^{-1}\rho(c_a^2 + c_s^2)}{2(1-\rho)} , \tag{9.2}$$

where $\rho \equiv \lambda/\mu$ is the traffic intensity. (We obtain (9.2) by multiplying (6.13) by ρ^2, which provides an asymptotically exact formula as $\rho \uparrow 1$ for any $GI/GI/1$ queue.

Part of the overall approximation error is due to using formula (9.2) for a $GI/GI/1$ queue. It is natural to ask about the range of possible mean-waiting-time values consistent with the four specified parameters; that is investigated for the $GI/M/1$ special case in Whitt (1984b,c) and Klincewicz and Whitt (1984). The range of possible values given the partial specification is quite large, e.g., the relative error could well be 100%, but for "typical" distributions, the range is not great, so that the relative error might be only 10%. However, touting much better accuracy, such as 1% relative error, for specific interarrival-time and service-time distributions is pointless because we can find different distributions that yield larger errors.

It is possible to improve (9.2), but we cannot escape the inevitable error caused by the partial characterization. A good refinement to approsimation (9.2), which makes EW smaller in some cases, was developed by Kraemer and Langenbach-Belz (1976). Possible refinements are discussed in Whitt (1983a, 1989b, 1993a) and references cited there.

Given approximation (9.2), the problem is to approximate the arrival and service processes at each queue in the network by the arrival and service processes in a $GI/GI/1$ queue partially characterized by the parameter four-tuple $(\lambda, c_a^2, \mu, c_s^2)$.

We start by treating the aggregate exogenous arrival process at each queue as the superposition of the single-class exogenous arrival processes at that queue. The exogenous arrival rate clearly should be the sum of the component single-class exogenous arrival rates at that queue. The variability function for the aggregate exogenous arrival process is more complicated and will be discussed later. The routing of customers within the network is

treated as Markovian: The probability $P_{i,j}$ of a customer going next to queue j after completing service at queue i is made equal to the long-run proportion of departing customers from queue i that are routed next to queue j. At each queue, the first two moments of the aggregate service-time distribution is just the weighted (by the arrival rates) average of the moments of the individual service-time distributions. The service-time SCV is defined in terms of the first two moments in the usual way: $c_s^2 + 1 \equiv E[V^2]/(E[V])^2$.

We act as if the service times do not need great care, and that often is the case. However, Example 9.6.1 illustrates how there can be significant dependence among successive service times and significant dependence between interarrival times and service times. In any specific application setting, it is good to have verification by simulation and measurement. We can verify both the final performance predictions and the variability characterizations of arrival and service processes.

It is straightforward accounting to produce aggregate exogenous arrival rates, Markovian routing probabilities at each queue and service-time distributions partially characterized by their first two moments. Indeed, the first-order deterministic rate parameters are exact in the specified procedure.

In engineering applications of queueing network analyzers, e.g., in the design of a manufacturing facility, usually most of the benefit is gained from the initial phase of the analysis. In the initial planning stages, the model formulation and accounting identify queues with unacceptably high traffic intensities (e.g., $\rho_i > 1$).

A second benefit that occurs before solving the model comes from having a model with an explicit quantification of variability. The form of the required model data focuses attention on variability. It indicates what should be measured. To build the queueing-network model, the engineers must look at process variability. When engineers attempt to measure and quantify the variability of arrival and service processes, they often discover opportunities to reduce that variability and make dramatic improvements in system performance.

Returning to the parametric-decomposition approximation, it remains to determine the SCV of the renewal arrival process approximating the arrival process at each queue. We can decompose the final approximation of c_a^2 for one such queue into two steps: (i) approximating the exogenous arrival process at each queue by a renewal process partially characterized by its rate and SCV, and (ii) approximating the net arrival process at the queues in the network by renewal arrival processes partially characterized by their rates and SCV's. (The exact rates of both the exogenous and aggregate arrival processes have already been determined.)

The second step involves developing an approximation for a *generalized Jackson network*, which is a single-class queueing network with Markovian routing, mutually independent renewal exogenous arrival processes and IID service times at the queues. For the generalized Jackson network considered here, the interarrival-time and service-time distributions are only partially characterized by their first two moments or, equivalently, by their means and SCV's. Dividing the overall approximation into two steps allows us to focus on the accuracy of each step separately.

9.9.2. Approximately Characterizing Arrival Processes

We now discuss ways to approximate a general arrival process with stationary interarrival time sequence $\{U_k : k \geq 1\}$ by a renewal process partially characterized by the mean λ^{-1} and SCV c_a^2 of an interarrival time. Following Whitt (1982a), we observe that there are two natural ways: In both ways, we let the arrival rate be specified exactly by letting $\lambda^{-1} = EU_1$. The *stationary-interval method* lets the SCV c_a^2 be the SCV of one interval U_1, i.e., we let

$$c_a^2 \approx c_{SI}^2 \equiv c_u^2 \equiv Var(U_1)/(EU_1)^2 \ , \tag{9.3}$$

as in (6.7). The *asymptotic method* lets c_a^2 be the scaled asymptotic variance

$$c_a^2 \approx c_{AM}^2 \equiv c_U^2 \equiv \lim_{n\to\infty} \frac{Var\, S_n^u}{n(EU_1)^2} \ , \tag{9.4}$$

where $S_n^u \equiv U_1 + \cdots + U_n$, $n \geq 1$, as in (6.11).

Under the regularity condition of uniform integrability, the asymptotic method in (9.4) is equivalent to c_U^2 being the dimensionless space-scaling constant in the CLT for S_n^u or the associated arrival counting process $A(t)$, i.e.,

$$(c_U^2 \lambda^{-2} n)^{-1/2}(S_n^u - \lambda^{-1} n) \Rightarrow N(0,1) \tag{9.5}$$

or, equivalently,

$$(\lambda c_U^2 t)^{-1/2}(A(t) - \lambda t) \Rightarrow N(0,1) \ ; \tag{9.6}$$

see Sections 7.3 and 13.8.

The stationary-interval method in (9.3) ignores any correlations among successive interarrival times. At first glance, it might appear that the stationary-interval method is *implied* by a renewal-process approximation, because there are *no correlations* in a renewal process, but that is not so. Even though the approximating process is to be viewed as a renewal process, it is important *not* to ignore the correlations in the arrival process being approximated if significant correlations are there.

In contrast, the asymptotic method includes *all the correlations* in the arrival process being approximated. From the Brownian heavy-traffic limits for general $G/G/1$ queues discussed in Section 9.6, we know that the asymptotic method is asymptotically correct in heavy traffic, using the mean waiting-time formula in (9.2). Thus, the heavy-traffic limit provides a very important reference point for these heuristic approximations.

However, in light traffic the long-run correlations among interarrival times obviously are not relevant, so that the stationary-interval method seems intuitively better in light traffic. Indeed, the stationary-interval method usually performs well in light traffic. To appreciate this discussion, it is important to realize that the two approximation procedures can both perform well in their preferred regimes, and yet c_{AM}^2 can be very very different from c_{SI}^2. (We will give examples below.) Thus neither procedure alone can always work well.

Thus, an effective approximation procedure needs to involve a compromise between the two basic approaches. As mentioned in Section 5.7, one possible approach is to interpolate between light-traffic and heavy-traffic limits, but we do not discuss that approach.

9.9.3. A Network Calculus

A parametric-decomposition algorithm for open queueing networks provides an algorithm for calculating the approximate arrival-process variability parameter c_a^2 at each queue in the network. That variability parameter will subsequently be used, together with the exact arrival rate, to approximately characterize an approximating renewal arrival process at that queue. The overall algorithm for calculating the arrival-process variability parameters can be based on a *network calculus* that transforms arrival-processe variability parameters for each of the basic network operations: superposition, splitting and departure (flow through a queue).

When the network is acyclic, the basic transformations can be applied sequentially, one at a time, but in general it is necessary to solve a system of equations in order to calculate the final variability parameters. Solving the equations becomes elementary if all the transformations are linear. Then the final algorithm involves solving a system of linear equations, with one equation for each queue. Hence there is motivation for developing linear approximations to characterize each transformation. The synthesis into a final system of linear equations is relatively straightforward; see Whitt (1983a, 1995); we will not discuss it here.

Here we will only discuss the basic transformations and the initial choice

of variability parameters. As mentioned earlier, we will focus on variability functions instead of variability parameters. Given a variability function $\{c_a^2(\rho) : 0 \leq \rho \leq 1\}$, we obtain a specific variability parameter when we specify the traffic intensity at the queue.

Superposition. Superposition applies first to the exogenous arrival process at each queue and then to the aggregate or net arrival process at each queue, including departures routed from other queues. Suppose that we have the superposition of m independent renewal counting processes $A^i(t)$ with rates λ_i and SCVs $c_{a,i}^2$. As indicated above, these parameters can be determined from the first two moments of an interarrival time. Alternatively, the parameters can be determined from a CLT for A^i of the form

$$[A^i(t) - \lambda_i t]/\sqrt{\lambda_i c_{a,i}^2} \Rightarrow N(0,1) \ . \tag{9.7}$$

Clearly the rate of the superposition process $A \equiv A^1 + \cdots A^m$ is $\lambda \equiv \lambda_1 + \cdots + \lambda_m$. It follows from Theorem 9.4.1 and Corollary 9.4.1 that the appropriate asymptotic-method approximation for c_a^2 is the weighted average of the component SCV's, i.e.,

$$c_{AM}^2 \equiv \sum_{i=1}^{m} (\lambda_i/\lambda) c_{a,i}^2 \ . \tag{9.8}$$

The stationary-interval method for superposition processes is more complicated, as can be seen from exact formulas in Section 4.1 of Whitt (1982a). However, by Theorem 9.8.1, for large m the superposition process behaves locally like a Poisson process, so that a large-m stationary-interval approximation is

$$c_{SI}^2 \approx 1 \ . \tag{9.9}$$

Notice that we have a demonstration of the inconsistency of the two basic approximation methods: For a superposition of m IID renewal processes, no matter how large is the interarrival-time SCV in a component arrival process, the superposition process approaches a Poisson process as $m \to \infty$. If the traffic intensity in a following queue is not too large, then the congestion at the queue is essentially the same as if the superposition arrival process were a Poisson process. On the other hand, for any fixed m, if the traffic intensity is high enough, the heavy-traffic limit is approximately correct. Since c_{AM}^2 can be arbitrarily large, the error from making the wrong choice can be arbitrarily large.

On the other hand, if $c_{AM}^2 \approx 1$, then the two basic methods are consistent and a Poisson-process approximation for the arrival process, which has $c_a^2 = 1$, is likely to perform well in many applications. However, if c_{AM}^2 is not near c_{SI}^2, then we can consider that a demonstration that the actual arrival process is not nearly a renewal process. Nevertheless, it may be possible to choose a variability parameter c_a^2 so that (9.2) is a reasonably good approximation for the mean waiting time.

The problem then is to find a compromise between the asymptotic method and the stationary-interval method that is appropriate for the queue. In general, that should depend upon the traffic intensity in the following queue. From the heavy-traffic limits in Section 9.3, it follows that the asymptotic method is asymptotically correct for the queue as $\rho \uparrow 1$, so that we should have $c_a^2(\rho) \to c_{AM}^2$ as $\rho \uparrow 1$. On the other hand, for very small ρ it is apparent that the stationary-interval method should be much better, so that we should have $c_a^2(\rho) \to c_{SI}^2$ as $\rho \downarrow 0$.

We can use Theorem 9.8.3 as a theoretical basis for a refined approximation. From Theorem 9.8.3, we know that, for superposition arrival processes with m component arrival processes, where $m \to \infty$, the asymptotic method is asymptotically correct for the queue as $\rho \to 1$ only if $m(1-\rho)^2 \to 0$. Thus, with superposition arrival processes, the weight on the asymptotic method should be approximately inversely proportional to $m(1-\rho)^2$.

In general, we want to treat superposition arrival processes where the component arrival processes have different rates. The number m has precise meaning in the expression $m(1-\rho)^2$ above only for identically distributed component processes. If one component process has a rate much larger than the sum of the rates of all other component processes, then the effective number should only be slightly larger than 1, regardless of m. However, it is not difficult to identify appropriate "equivalent numbers" of component processes that allow for unequal rates.

The considerations above lead to generalizations of the approximation used in the queueing network analyzer (QNA) software tool; see Whitt (1983a, 1995), Albin (1984) and Segal and Whitt (1989).

Specifically, an approximating variability function $c_a^2(\rho)$ for the superposition arrival process is

$$\begin{aligned} c_a^2(\rho) &\approx w c_{AM}^2 + (1-w) c_{SI}^2 \\ &\approx w \left(\sum_{i=1}^{m} (\lambda_i/\lambda) c_{a,i}^2(\rho) \right) + (1-w) \ , \end{aligned} \tag{9.10}$$

where

$$w \equiv w(\rho, \nu) \equiv [1 + 4(1-\rho)^2(\nu - 1)]^{-1} \tag{9.11}$$

with

$$\nu \equiv \left[\sum_{i=1}^{m} (\lambda_i/\lambda)^2 \right]^{-1} . \tag{9.12}$$

The parameter ν in (9.12) is the "equivalent number" of component arrival streams, taking account of unequal rates. When $m = 1$, $\nu = 1$ and $c_a^2(\rho) = c_{a,1}^2(\rho)$. In (9.10) we use the approximation $c_{SI}^2 \equiv c_{a,i}^2(0) \approx 1$ motivated by Theorem 9.8.1. Notice that w as a function of ρ and ν is roughly consistent with the scaling in (8.9) in Theorem 9.8.3: The complex limit occurs as $\nu(1-\rho)^2$ converges to a nondegenerate limit.

Splitting. When the routing is Markovian and we start with a renewal process, the split processes are also renewal processes, so that $c_{AM}^2 = c_{SI}^2$. If a renewal arrival process with interarrival times having mean λ^{-1} and SCV c_a^2 is split into m streams, with the probability being p_i of each point being assigned to the i^{th} split stream, then the mean and SCV of the interarrival time in the i^{th} split stream are

$$\lambda_i^{-1} = (\lambda p_i)^{-1} \quad \text{and} \quad c_{a,i}^2 = p_i c_a^2 + 1 - p_i \ , \tag{9.13}$$

as can be deduced from Theorem 9.5.1.

We now want to extend the splitting formula to independent splitting from more general non-renewal processes. Now the original arrival process is partially characterized by its arrival state λ and its variability function $\{c_a^2(\rho) : 0 \le \rho \le 1\}$, where ρ is the traffic intensity at the following queue. A natural generalization of (9.13) is

$$\lambda_i = \lambda p_i \quad \text{and} \quad c_{a,i}^2(\rho) = p_i c_a^2(\rho) + 1 - p_i \tag{9.14}$$

for $0 \le \rho \le 1$.

However, formulas (9.13) and (9.14) can perform poorly when the routing is not actually Markovian. Discussions of alternative approximations associated with non-Markovian routing appear in Bitran and Tirupati (1988) and Whitt (1988, 1994, 1995). In particular, when there are multiple classes with each class having its own deterministic routing, we can use the separation of time scales to deduce that the single-class departure process is closely related to the single-class arrival process: With many classes, the queue operates in a shorter time scale than the flow for one customer class. Then the

customer sojourn times, being relatively short compared to the single-class interarrival times, tend to make the single-class departure process differ little from the single-class arrival process.

Suppose that there are m single-class arrival processes with variability functions $\{c_{a,i}^2(\rho) : 0 \le \rho \le 1\}$ for $1 \le i \le m$. Let $\{c_{d,i}^2(\rho) : 0 \le \rho \le 1\}$ be the associated variability functions for the single-class departure processes from that queue. The separation of time scales suggests that we should have

$$c_{d,i}^2(\rho) \approx c_{a,i}^2(\rho), \quad 0 \le \rho \le 1 \ , \tag{9.15}$$

The approximation (9.15) treats departure and splitting together in one step.

Departures. The stationary interval between departures in a $GI/GI/1$ queue partially characterized by the parameter four-tuple $(\lambda, c_a^2, \mu, c_s^2)$ has mean λ^{-1} and SCV

$$c_d^2 = c_a^2 + 2\rho^2 c_s^2 - 2\rho(1-\rho)\mu EW \ . \tag{9.16}$$

Hence we can use (9.2) to produce an approximation for the stationary-interval SCV of a departure process,

$$c_{SI}^2 \approx \rho^2 c_s^2 + (1-\rho^2)c_a^2 \ ; \tag{9.17}$$

see Whitt (1984d). However, except for the $M/M/1$ queue, the departure process is not a renewal process. Hence there are correlations among successive interdeparture times that are not captured by approximation (9.17). Nevertheless, simulation experiments indicate that approximation (9.17) often performs remarkably well. For example, simulations indicate that approximations (9.17) and (9.2) together work well to determine the best order for queues in series (to minimize the mean steady-state sojourn time, given a fixed arrival process); see Whitt (1985b) and Suresh and Whitt (1990b).

As noted in Remark 5.3.1, for $0 < \rho < 1$, the departure process obeys the same CLT as the arrival process. Thus the asymptotic-method approximation for the departure process is

$$c_{AM}^2 = c_a^2 \ . \tag{9.18}$$

To highlight the difference between (9.17) and (9.18), consider two queues in series – the $GI/GI/1 \to /GI/1$ model. Let ρ_i be the traffic intensity, $c_{a,i}^2$ the interarrival-time SCV and $c_{s,i}^2$ the service-time SCV at queue i for

$i = 1, 2$. First, it is evident that the departure process from queue 1 approaches the service process there as $\rho_1 \uparrow 1$. Consistent with that property, $c_{SI}^2 \to c_{s,1}^2$ as $\rho_1 \uparrow 1$ by (9.17). On the other hand, the asymptotic-method approximation is asymptotically correct for the arrival process at the second queue as $\rho_2 \uparrow 1$. Hence $c_{AM}^2 = c_{a,1}^2$ is asymptotically correct for $c_{a,2}^2$ as $\rho_2 \uparrow 1$ for fixed ρ_1.

Now, turning to the variability functions, a candidate approximation consistent with the reference point above is

$$c_{a,2}^2(\rho_2) = c_{d,1}^2(\rho_1, \rho_2) = \alpha(\rho_1, \rho_2)c_{s,1}^2 + (1 - \alpha(\rho_1, \rho_2))c_{a,1}^2(\rho_2) \ , \qquad (9.19)$$

where $\alpha(\rho_1, \rho_2) \uparrow 1$ as $\rho_1 \uparrow 1$ and $\alpha(\rho_1, \rho_2) \downarrow 0$ as $\rho_2 \uparrow 1$. A specific candidate that agrees with (9.17) unless $\rho_2 > \rho_1$ is

$$\alpha(\rho_1, \rho_2) = \rho_1^2 min\{1, (1 - \rho_2)^2/(1 - \rho_1)^2\} \ , \qquad (9.20)$$

but further study is needed.

Example 9.9.1. *The heavy-traffic bottleneck phenomenon.* The purpose of this example is to demonstrate the need for variability functions instead of variability parameters to partially characterize arrival processes in parametric-decomposition approximations. We consider a large number n of queue in series, all with relatively low traffic intensity ρ_1, followed by a $(n+1)^{\text{st}}$ queue with high traffic intensity ρ_{n+1}.

To be concrete, we consider a $GI/M/1 \to /M/1 \to \cdots \to /M/1$ model with a rate-1 renewal arrival process partially characterized by its SCV $c_{a,1}^2$. The service-time distributions are all exponential, so that $c_{s,i}^2 = 1$ for all i. The mean service time and traffic intensity at each of the first n queues is ρ_1, while the traffic intensity at the final $(n+1)^{\text{st}}$ queue is ρ_{n+1}.

It is known that as n increases, the stationary departure process from the n^{th} queue approaches a Poisson process; see Mountford and Prabhakar (1995), Mairesse and Prabhakar (2000) and references cited there. Consistent with that limit, the stationary-interval approximation in (9.17) for the SCV $c_{a,n+1}^2$ satisfies

$$c_{SI,n+1}^2 = (1 - \rho_1^2)^n c_a^2 + (1 - (1 - \rho_1^2)^n) \to 1 \qquad (9.21)$$

as $n \to \infty$. On the other hand, for any fixed ρ_1, the final $(n+1)^{\text{st}}$ queue has a heavy-traffic limit that depends on the first n queues only through the exogenous arrival rate 1 and the SCV $c_{a,1}^2$.

We now describe a simulation experiment conducted by Suresh and Whitt (1990a) to show that this heavy-traffic bottleneck phenomenon is of

practical significance. To consider "typical" values, they let $n = 8$, $\rho_1 = 0.6$ and $\rho_9 = 0.9$. (The initial traffic intensity is not too low, while the final traffic intensity is not too high.) Two renewal arrival processes are considered: hyperexponential interarrival times (mixtures of two exponential distributions) with $c_{a,1}^2 = 8.0$ and deterministic interarrival times with $c_{a,1}^2 = 0.0$, representing high and low variability.

We compare simulation estimates of the mean steady-state waiting times with three approximations. In all three approximations, the approximation formula is

$$EW \approx \frac{\rho^2(c_a^2 + 1)}{2(1 - \rho)} , \tag{9.22}$$

which is obtained from (9.2) by letting $\mu^{-1} = \rho$ and $c_s^2 = 1$. The three approximations differ in their choice of the arrival-process variability parameter c_a^2: The asymptotic-method (or heavy-traffic) approximation lets $c_a^2 = c_{a,1}^2$; the stationary-interval approximation lets $c_a^2 = c_{SI,n+1}^2$; the $M/M/1$ approximation lets $c_a^2 = 1$. The SI approximation yields $c_{a,9}^2 = 1.20$ and $c_{a,9}^2 = 0.97$ in the two cases.

Table 9.2 shows the results of the simulation experiment. From Table 9.2, we see that the asymptotic-method approximation is far more accurate than the other two approximations at the final queue 9, while the other two approximations are far more accurate at the previous queue 8 with lower traffic intensity. The appropriate variability parameter for the arrival process clearly depends on the traffic intensity at the final queue.

Consistent with the different approximations at the queues, the measured variability parameters differ. The stationary interarrival time at queue 9 has an SCV close to 1, while the estimated asymptotic variability parameter $c_{U,9}^2$ is close to $c_{a,1}^2$. Just as with superposition arrival processes (see Albin (1982)), the individial lag-k correlations are small; $c_{U,9}^2$ differs from $c_{u,9}^2$ because of the cumulative effect of many small correlations.

Just as in examples with superposition arrival processes, the heavy-traffic bottleneck phenomenon illustrates the need for variability functions. The heavy-traffic bottleneck phenomenon also illustrates that there can be long-range variability effects in networks. High or low variability in an exogenous arrival process can be unseen (can have little congestion impact) in some queues and then suddenly appear at a later queue with a much higher traffic intensity. The reason is that different levels of variability can exist at different time scales. The arrival process to the final queue in this example looks like a Poisson process in a small time scale, but looks like the exogenous arrival process in a long time scale.

		High variability $c_{a,1}^2 = 8.0$	Low variability $c_{a,1}^2 = 0.0$
Queue 9 $\rho_9 = 0.9$	Simulation estimate	30.1 ± 5.1	5.03 ± 0.22
	asymptotic-method approximation	36.5	4.05
	stationary-interval approximation	8.9	8.0
	M/M/1 approximation	8.1	8.1
Queue 8 $\rho_8 = 0.6$	Simulation estimate	1.42 ± 0.07	0.775 ± 0.013
	asymptotic-method approximation	4.05	0.45
	stationary-interval approximation	1.04	0.88
	M/M/1 approximation	0.90	0.90

Table 9.2: A comparison of approximations with simulation estimates of the mean steady-state waiting times at queue 9 and 8 in the network of nine queues in series.

9.9.4. Exogenous Arrival Processes

In applications of any method for analyzing the performance of queueing networks, it is necessary to specify the exogenous arrival processes. With the parametric-decomposition approximation, it is necessary to obtain initial variability functions chacterizing the exogenous arrival processes. If the exogenous arrival processes are actually renewal processes, then there is no difficulty: then we can simply let the variability function $c_a^2(\rho)$ be the SCV of an interarrival time for all traffic intensities ρ.

However, experience indicates that, in practice (as opposed to in models), an arrival process that fails to be nearly a Poisson process also fails to be nearly a renewal process. Indeed, exogenous arrival processes often fail to be renewal processes, so that it is necessary to take care in characterizing

the variability of these exogenous arrival processes. Hence, instead of the route vector in (9.1), the model data for that customer class should be of the form

$$(1, 2, \{c_{a,0}^2(\rho) : 0 \leq \rho \leq 1\}; 2, 1, 0; 3, 1, 1; 2, 5, 1) \ . \tag{9.23}$$

With variability functions, then, we should be prepared to specify the variability functions of the exogenous arrival processes. Following Whitt (1981, 1983c) and Section 3 of Whitt (1995), we suggest fitting variability parameters indirectly by observing the congestion produced by this arrival process in a test queue. This can be done either through analytical formulas (if the arrival process is specified as a tractable mathematical model) or through simulation (if the arrival process is specified either as a mathematical model or via direct system measurements).

For example, we can use approximation formula (9.2). We might consider an exponential service-time distribution, which makes $c_s^2 = 1$. We then think of the queue as a $GI/M/1$ queue, but since the arrival process may not actually be a renewal process, we allow the variability parameter to depend on the traffic intensity. We estimate the mean waiting time as a function of ρ using the arrival process to be characterized. For each value of ρ, we let the variability function $c_a^2(\rho)$ assume the value c_a^2 that makes formula (9.2) match the observed mean waiting time.

This indirect procedure is illustrated by applying it to *irregular periodic deterministic arrival processes* in Section 4 of Whitt (1995). A simple example has successive interarrival times 3/2, 1/2, 3/2, 1/2, ... Consistent with intuition, for irregular periodic deterministic arrival processes, $c_a^2(\rho) = 0$ for all sufficiently small ρ and $c_a^2(\rho) \to 0$ as $\rho \uparrow 1$, but $c_a^2(\rho)$ can be arbitrarily large for intermediate values of ρ.

This indirect estimation procedure can also be used to refine parametric-decomposition approximations for the variability functions partially characterizing the internal flows in the network. Through simulations and measurements, we can appropriate variability functions that lead to accurate performance predictions for the internal flows, just as for the exogenous arrival processes. See Fendick and Whitt (1989) and Whitt (1995) for further discussion.

9.9.5. Concluding Remarks

We conclude this section with two remarks.

Remark 9.9.1. *Heavy-traffic limits for queueing networks.* An alternative to the parametric-decomposition approximation is an approximation based

directly on a heavy-traffic limit for the queueing network. The heavy-traffic limit ideally would be for the original multiclass queueing network, but it could be for the single-class Jackson network constructed in the first phase of the procedure described above. Heavy-traffic limits for the single-class generalized Jackson network are developed in Chapter 14. A specific algorithm based on the heavy-traffic limit is the QNET algorithm of Dai (1990), Dai and Harrison (1991, 1992), Harrison and Nguyen (1990) and Dai, Yeh and Zhou (1997). A direct heavy-traffic algorithm is an attractive alternative, but the limit process is usually complicated. The computational complexity of the QNET algorithm grows rapidly as the number of nodes increases.

When considering heavy-traffic limits for queueing networks, it is important to recognize that there is more than one way to take the heavy-traffic limit. With a queueing network, there is more than one traffic intensity: There is a traffic intensity at each queue. The standard limiting procedure involves *balanced loading*, in which all the traffic intensities approach 1 together; i.e., if ρ_i is the traffic intensity at queue i, then $\rho_i \uparrow 1$ for all i with $(1-\rho_i)/(1-\rho_1) \to c_i$, $0 < c_i < \infty$.

However, the bottleneck view, stemming from consideration of a fixed network with one traffic intensity larger than the others, has one traffic intensity approach 1 faster than the others. If the traffic intensity at one queue approaches 1 faster than the traffic intensities at the other queues, then we see a nondegenerate limit for the scaled queue-length process only at the bottleneck queue. With this form of heavy-traffic limit, one queue dominates. Just as in the heavy-traffic bottleneck phenomenon, the heavy-traffic approximation is equivalent to the heavy-traffic limit in which all the service times at the other queues are reduced to zero, and the other queues act as instantaneous switches.

A more general heavy-traffic approximation for a network of queues isthe sequential-bottleneck decomposition method proposed by Reiman (1990a) and Dai, Nguyen and Reiman (1994). It is a heirarchical procedure similar to the one proposed for the priority queue in Section 5.10. The sequential-bottleneck procedure decomposes the network into groups of one or more queues with similar traffic intensities. Then heavy-traffic approximations are developed for the groups separately, starting with the group with highest traffic intensities. When analyzing a subnetwork associated with a group of queues, the remaining queues are divided into two sets, those with larger traffic intensities and those with smaller traffic intensities. Queues with smaller traffic intensities are treated as if their service times are zero, so they act as instantaneous switches. Queues with larger traffic intensities are treated as if they are overloaded, which turns them into sinks for flows

into them and exogenous sources for flows out of them. Then the QNET approximation is applied to each subgroup. If the subgroups only contain a single queue, then we can apply the simple single-queue heavy-traffic approximation. The single-queue case was proposed by Reiman (1990a).

The sequential-bottleneck approximation is appealing, but note that it offers no way to achieve the needed non-heavy-traffic approximation at a queue with a superposition arrival process having many components. At first glance, the single-queue sequential-bottleneck approximation seems to perform well on Example 9.9.1: It produces the heavy-traffic approximation at the final queue with high traffic intensity, which is pretty good. However, the heavy-traffic approximation at the final queue would not be good if we lowered the traffic intensity of the final queue from 0.9 to 0.61, where it still is greater than all other traffic intensities. It still remains to develop an approximation for the special $GI/M/1 \rightarrow \cdots \rightarrow /M/1$ model with nine queues in series that can be effective for all possible traffic-intensity vectors. ∎

Remark 9.9.2. *Closed queueing networks.* For many applications it is natural to use closed queueing network models, which have fixed customer populations, instead of open queueing network models. There are convenient algorithms for a large class of Markovian closed queueing network models, but non-Markovian closed queueing network models tend to be intractable.

Approximations for non-Markovian open queueing networks can be applied via the *fixed-population-mean* (FPM) method: The steady-state performance of the closed queueing network is approximated by the steady-state performance of an associated open network in which the mean population in the open network is set equal to the specified population in the closed network; see Whitt (1984c). A search algorithm identifies the exogenous arrival rate in the open model producing the target mean. (A more complicated search algorithm is required if there are multiple customer classes with specified populations.) The FPM method provides good approximations when the population is not too small.

The FPM method can explain seemingly anomalous behavior in non-Markovian closed queueing networks: If the variability of the service-time distribution increases at one queue, then it is possible for the mean queue length at that queue to decrease. Indeed that phenomenon routinely occurs at a bottleneck queue; see Bondi and Whitt (1986). That occurs because the bottleneck queue tends to act as an exogenous source for the rest of the network. Thus increased variability at the bottleneck queue is likely to cause

greater congestion in the rest of the network. Since the total population is fixed, the mean queue length at the bottleneck queue is likely to go down. ■

To summarize, parametric-decomposition approximations for queueing networks can be great aids in performance analysis. And heavy-traffic limits can help improve the performance of these algorithms. However, at the present time there is no one algorithm that works well on all examples. Nevertheless, there is sufficient understanding and there are sufficient tools to make effective algorithms for many specific classes of applications.