Chapter 10

Multi-Server Queues

10.1. Introduction

We continue applying the continuous-mapping approach to establish heavy-traffic stochastic-process limits for stochastic processes of interest in standard queueing models, but now we consider multi-server queueing models. There are two principal cases: first, when there is a moderate number of servers and, second, when there is a large number of servers. The first case commonly occurs in manufacturing systems, which may have workstations containing several machines. The second case commonly occurs in call centers, which may have agent groups containing hundreds of agents. These two cases are sufficiently different to warrant different methods.

We start in Section 10.2 by considering the first case of a fixed finite number of servers. We show that the heavy-traffic behavior for a fixed finite number of servers is essentially the same as a single-server system (with an obvious scale adjustment to account for the multiple servers).

We consider the second case of a large number of servers in the remaining sections. The natural approximation for a large number of servers is an infinite number, provided that we find appropriate measures of congestion. In Section 10.3 we consider the case of infinitely many servers.

In Section 10.4 we consider the case in which the number of servers increases along with the traffic intensity in the heavy-traffic limit, so that the probability of delay converges to a nondegenerate limit. In Section 10.4 we also consider multi-server loss systems, without any extra waiting room. Paralleling Section 5.8, we show how heavy-traffic limits can be used to determine the simulation run lengths required to estimate, with prescribed precision, blocking probabilities in loss models with a large number of servers. Interestingly, the story is quite different from the single-server

case in Section 5.8.

10.2. Queues with Multiple Servers

In this section we consider a queue with a fixed finite number, m, of servers. We should anticipate that the heavy-traffic behavior of a multi-server queue is essentially the same as the heavy-traffic behavior of a single-server queue with a superposition arrival process, treated in Section 9.4: Now we have a multi-channel output process instead of a multi-channel input process. Indeed, that is the case, but providing a proper demonstration is somewhat more difficult.

10.2.1. A Queue with Autonomous Service

One approach to this problem, used by Borovkov (1965) and Iglehart and Whitt (1970a, b), is to relate the standard multi-server queue to another model that is easier to analyze. Another model that is easier to analyze is a *queue with autonomous service*, in which servers are not shut off when they become idle. Associated with each of the m servers is a sequence of potential service times. If a server faces continued demand, then the actual service times coincide with these potential service times, but if there is no demand for service, then these potential service times are ignored and there is no actual service and no departure. New arrivals are assigned to the server that can complete their service first. After a server has been working in the absence of demand, the next demand will in general fall in the middle of a service time. The service time of that customer arriving after the server has been idle is the remaining portion of the potential service time in process at that time.

The queue with autonomous service is of some interest in its own right, but it was introduced primarily as a device to treat the standard multi-server queueing model. In the standard model, customers are assigned in order of arrival to the first available server, with some unspecified procedure to break ties. The queue with autonomous service is easy to analyze because the reflection map can be applied directly.

The service process in the queue with autonomous service is just like a superposition arrival process. Let $\{V_{i,k} : k \geq 1\}$ be the sequence of potential service times for server i for $1 \leq i \leq m$. Let the associated partial sums be

$$S^v_{i,k} \equiv V_{i,1} + \cdots + V_{i,m}, \quad k \geq 1, \tag{2.1}$$

and let $S^v_{i,0} \equiv 0$. Let N_i be the associated counting process, defined as in (2.9) by

$$N_i(t) \equiv max\{k \geq 0 : S^v_{i,k} \leq t\}, \quad t \geq 0. \tag{2.2}$$

Let N be the superposition process defined by

$$N(t) \equiv N_1(t) + \cdots + N_m(t), \quad t \geq 0 \tag{2.3}$$

and let S^v be the inverse partial-sum process associated with N, defined by

$$S^v_k \equiv inf\{t \geq 0 : N(t) \geq k\}, \quad k \geq 0, \tag{2.4}$$

with $S^v_0 \equiv 0$. Let

$$V_k \equiv S^v_k - S^v_{k-1}, \quad k \geq 1 \ . \tag{2.5}$$

Note that the potential-service processes S^v_i, N_i, N, S^v and V are related the same way that superposition-arrival-process processes S^u_i, A_i, A, S^u and U are related in (4.1)–(4.4).

For simplicity, let the queue start out empty. Let $\{A(t) : t \geq 0\}$ be the arrival counting process and let $Q^a(t)$ be the queue length, with the superscript "a" denoting autonomous service. The queue with autonomous service is defined so that the queue-length process is directly the reflection of the net-input process $A - N$, i.e.,

$$Q^a(t) = \phi(A - N)(t), \quad t \geq 0 \ , \tag{2.6}$$

where ϕ is the one-sided reflection map in (2.5) of Section 9.2. By focusing on the m-server queue with autonomous service, the queue-length process becomes a special case of the fluid-queue model in Chapter 8 in which the cumulative-input and available-processing processes are integer-valued. The arrival process A is the cumulative-input process, while the service-time counting process N is the available-processing process.

Remark 10.2.1. *The case of IID exponential service times.* To better understand the queue with autonomous service, consider the special case in which the potential service times come from m independent sequences of IID exponential random variables with mean 1, independent of the arrival process. Then the queue with autonomous service is not equivalent to the $G/M/m$ queue, but is instead equivalent to the $G/M/1$ queue with the service rate m. In general, the $G/M/1$ queue with service rate m is quite different from the associated $G/M/m$ queue, having identical arrival process and m servers each with rate 1. However, for fixed m, in heavy traffic they behave essentially the same. Indeed, that is established as a special case of Theorem 10.2.2 below. ∎

Now let us consider a sequence of these queueing models with autonomous service indexed by n. Paralleling (4.5), define associated random elements of $D \equiv D([0,\infty),\mathbb{R})$ by letting

$$\begin{aligned}
\mathbf{S}^v_{n,i}(t) &\equiv c_n^{-1}[S^v_{n,i,\lfloor nt\rfloor} - \mu_{n,i}^{-1}nt], \\
\mathbf{N}_{n,i}(t) &\equiv c_n^{-1}[N_{n,i}(nt) - \mu_{n,i}nt], \\
\mathbf{N}_n(t) &\equiv c_n^{-1}[N_n(nt) - \mu_n nt], \\
\mathbf{S}^v_n(t) &\equiv c_n^{-1}[S^v_{n,\lfloor nt\rfloor} - \mu_n^{-1}nt] \ , \qquad (2.7)
\end{aligned}$$

where $\mu_n \equiv \mu_{n,1} + \cdots + \mu_{n,m}$. A limit for $(\mathbf{S}^v_{n,1},\ldots,\mathbf{S}^v_{n,m})$ implies an associated limit for $(\mathbf{S}^v_{n,1},\ldots,\mathbf{S}^v_{n,m},\mathbf{N}_{n,1},\ldots\mathbf{N}_{n,m},\mathbf{N}_n,\mathbf{S}^v_n)$ by Theorem 9.4.1.

Now define additional random elements of D associated with the arrival and queue-length processes A and Q^a by

$$\begin{aligned}
\mathbf{A}_n(t) &\equiv c_n^{-1}[A_n(nt) - \lambda_n nt], \\
\mathbf{Q}^a_n(t) &\equiv c_n^{-1}Q^a_n(nt), \quad t \ge 0 \ . \qquad (2.8)
\end{aligned}$$

The following heavy-traffic limit parallels Theorem 9.4.2, and is proved in the same way.

Theorem 10.2.1. (heavy-traffic limit for the m-server queue with autonomous service) *Suppose that*

$$(\mathbf{A}_n, \mathbf{S}^v_{n,1},\ldots\mathbf{S}^v_{n,m}) \Rightarrow (\mathbf{A}, \mathbf{S}^v_1,\ldots,\mathbf{S}^v_m) \quad \textit{in} \quad (D^{1+m}, WM_1) \qquad (2.9)$$

for $\mathbf{A}_n$ *in* (2.8) *and* $\mathbf{S}^v_{n,i}$ *in* (2.7). *Suppose that, for* $1 \le i \le m$,

$$P(\mathbf{A}(0) = 0) = P(\mathbf{S}^v_i = 0) = 1 \ , \qquad (2.10)$$

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

$$\eta_n \equiv n(\lambda_n - \mu_n)/c_n \to \eta \quad \textit{as} \quad n \to \infty \qquad (2.11)$$

for λ_n *in* (2.8) *and* $\mu_n \equiv \mu_{n,1} + \cdots + \mu_{n,m}$. *Suppose that*

$$P(Disc(\mathbf{S}^v_i \circ \mu_i\mathbf{e}) \cap Disc(\mathbf{S}^v_j \circ \mu_j\mathbf{e}) = \phi) = 1 \qquad (2.12)$$

and

$$P(Disc(\mathbf{S}^v_i \circ \mu_i\mathbf{e}) \cap Disc(\mathbf{A}) = \phi) = 1 \qquad (2.13)$$

for all i,j *with* $1 \le i,j \le m$, $i \ne j$. *Then*

$$(\mathbf{A}_n, \mathbf{N}_n, \mathbf{Q}^a_n) \Rightarrow (\mathbf{A}, \mathbf{N}, \mathbf{Q}) \quad \textit{in} \quad (D^3, WM_1) \ ,$$

where

$$\mathbf{N} = \mathbf{N}_1 + \cdots + \mathbf{N}_m = -\sum_{i=1}^{m} \mu_i \mathbf{S}_i^v \circ \mu_i \mathbf{e} \tag{2.14}$$

and

$$\mathbf{Q} = \phi(\mathbf{A} - \mathbf{N} + \eta \mathbf{e}) \ . \tag{2.15}$$

Remark 10.2.2. *Resource Pooling with IID Lévy processes.* Just as in Remark 9.4.1, if the m limit processes $\mathbf{S}_1^v, \ldots, \mathbf{S}_m^v$ are IID Lévy processes, then the limit processes associated with the superposition process are deterministic time-scalings of the limit process associated with a single server, i.e.,

$$\mathbf{N} \stackrel{\mathrm{d}}{=} \mathbf{N}_1 \circ m\mathbf{e} \quad \text{and} \quad \mathbf{S}^v \stackrel{\mathrm{d}}{=} m^{-1}\mathbf{S}_1^v \ .$$

Then the m servers act as a "single super server" and we say that there is *resource pooling*. We discuss resource pooling with other queue disciplines in Remark 10.2.4 below. ∎

10.2.2. The Standard m-Server Model

We now want to consider the standard m-server queue, in which customers wait in a single queue and are assigned in order of arrival to the first available server, with some unspecified procedure to break ties. We now want to show that the queue-length process in the standard m-server queue has the same heavy-traffic limit.

Iglehart and Whitt (1970a, b) showed that the scaled queue-length processes in the two systems are asymptotically equivalent under regularity conditions. As before, let $\mathbf{Q}_n^a$ denote the scaled queue-length process in the m-server queue with autononous service, just considered, and let $\mathbf{Q}_n$ denote the scaled queue-length process in the standard system, with the same scaling as in (2.8). The next result follows from Theorem 10.2.1 and the reasoning on pages 159–162 in Iglehart and Whitt (1970a).

Theorem 10.2.2. (asymptotic equivalence with the standard m-server queue) *In addition to the assumptions of Theorem* 10.2.1, *suppose that* $\{V_{i,k}^v : k \geq 1\}$, $1 \leq i \leq m$, *are* m *independent sequences of IID random variables, independent of the arrival process. If, in addition,*

$$P(\mathbf{S}_i^v \in C) = 1 \tag{2.16}$$

for each i, $1 \leq i \leq m$, *then there exist versions* $\mathbf{Q}_n$ *of the scaled standard queue-length process so that*

$$\|\mathbf{Q}_n - \mathbf{Q}_n^a\|_t \to 0 \quad w.p.1 \quad as \quad n \to \infty \ , \tag{2.17}$$

for all $t > 0$, *so that*

$$\mathbf{Q}_n \Rightarrow \mathbf{Q} \quad in \quad (D, M_1) \ ,$$

where $\mathbf{Q}$ *is as in* (2.15) .

Theorem 10.2.2 treats only one process in the standard multi-server queue. Under the assumptions of Theorem 10.2.2, heavy-traffic limits can also be obtained for other processes besides the queue-length process, as shown by Iglehart and Whitt (1970a, b).

Condition (2.16) requires the service-time limit processes $\mathbf{S}_i^v$ to have continuous paths. The asymptotic-equivalence argument in Iglehart and Whitt (1970a) does *not* apply if these processes can have jumps, because the difference is bounded above by the largest individual service time encountered, appropriately scaled. When the limit process has continuous sample paths, we can apply the maximum-jump function to conclude that this scaled maximum serice time is asymptotically negligible, but when the limit process has discontinuous sample paths, we cannot draw that conclusion.

Under the conditions of Theorem 10.2.2, the arrival-process limit process $\mathbf{A}$ can have discontinuous sample paths, so that in general we still need the M_1 topology to express the limit. The arrival process can force nonstandard scaling, but that may make the limit processes $\mathbf{S}_i^v$ be degenerate (zero processes), because the possibilities for the limit processes $\mathbf{S}_i^v$ are limited. Since the potential service times must come from sequences of IID random variables, to have the limit processes $\mathbf{S}_i^v$ be nondegenerate with continuous sample paths in Theorem 10.2.2, we are effectively restricted to the case in which $c_n = n^{1/2}$ and $\mathbf{S}_i^v$ is Brownian motion.

We can actually eliminate all the extra regularity conditions in Theorem 10.2.2 by using a different argument. Specifically, we can exploit bounds established by Chen and Shanthikumar (1994) to obtain convergence of the scaled standard queue-length processes under the assumptions of Theorem 10.2.1. Their argument actually applies to networks of multi-server queues.

Theorem 10.2.3. (heavy-traffic limit for the standard m-server queue) *Under the assumptions of Theorem* 10.2.1,

$$\mathbf{Q}_n \Rightarrow \mathbf{Q} \quad in \quad (D, M_1) \ ,$$

where $\mathbf{Q}_n$ *is the standard queue-length process scaled as in* (2.8) *and* $\mathbf{Q}$ *is in* (2.15) .

Proof. We use extremal properties of the regulator map ψ_L defined in equations (2.6)–(2.10) of Section 5.2: Suppose that x, y and z are three functions in D satisfying $z = x + y \geq 0$. If y is nondecreasing, $y(0) = 0$ and y increases only when $z(t) \leq b$ for some positive constant b, then

$$\psi_L(x)(t) \leq y(t) \leq \psi_L(x - b)(t) \quad \text{for all} \quad t \geq 0 \ . \tag{2.18}$$

The lower bound is proved in Section 14.2 – see Theorem 14.2.1 – and the upper bound is proved in the same way (also see Theorem 14.2.3), as shown by Chen and Shanthikumar (1994).

We now proceed to treat the queue-length process just as we did for the single-server queue in Theorem 9.3.4, allowing for the fact that we now have m servers: Paralleling (3.23), we obtain

$$\mathbf{Q}_n = \mathbf{X}_n + \mathbf{Y}_n \ , \tag{2.19}$$

where

$$\mathbf{X}_n = \mathbf{A}_n - (\sum_{i=1}^{m} \mathbf{N}_n^i \circ \hat{\mathbf{B}}_n^i) + \eta_n \mathbf{e} \ , \tag{2.20}$$

$$\hat{\mathbf{B}}_n^i(t) \equiv n^{-1} B_n^i(nt), \quad t \geq 0, \tag{2.21}$$

$$\mathbf{Y}_n \equiv \sum_{i=1}^{m} \mu_n^i \mathbf{Y}_n^i \ , \tag{2.22}$$

$$\mathbf{Y}_n^i(t) \equiv c_n^{-1}(nt - B_n^i(nt)), \quad t \geq 0 \ , \tag{2.23}$$

and $B_n^i(t)$ is the cumulative busy time of server i in the interval $[0, t]$ in model n. Note that $\mathbf{Y}_n$ increases only when $\mathbf{Q}_n(t)$ is less than or equal to m/c_n. Thus, by (2.18),

$$\psi_L(\mathbf{X}_n) \leq \mathbf{Y}_n \leq \psi_L(\mathbf{X}_n - m/c_n) \ . \tag{2.24}$$

Since ψ_L is a Lipschitz map in the uniform norm, by Lemma 13.4.1,

$$\|\mathbf{Y}_n - \psi_L(\mathbf{X}_n)\|_t \leq m/c_n \to 0 \quad \text{as} \quad n \to \infty$$

for each $t \geq 0$. Hence,

$$\|\mathbf{Q}_n - \phi(\mathbf{X}_n)\|_t \leq 2m/c_n \to 0 \quad \text{as} \quad n \to \infty$$

for each $t \geq 0$. Since ϕ is continuous as a map from (D, M_1) to itself, it suffices to show that $\mathbf{X}_n \Rightarrow \mathbf{X}$ in (D, M_1). (We use the convergence-together theorem, Theorem 11.4.7.)

The key to establishing the limit $\mathbf{X}_n \Rightarrow \mathbf{X}$ is to show that

$$(\hat{\mathbf{B}}_n^1, \ldots, \hat{\mathbf{B}}_n^m) \Rightarrow (\mathbf{e}, \ldots, \mathbf{e}) \tag{2.25}$$

for $\hat{\mathbf{B}}_n^i$ in (2.21). To establish the limit in (2.25), we exploit the compactness approach in Section 11.6: Specifically, we can apply Theorems 11.6.2, 11.6.3 and 11.6.7 after observing that $\hat{\mathbf{B}}_n^i$ is uniformly Lipschitz: For $0 < t_1 < t_2$,

$$|\hat{\mathbf{B}}_n^i(t_2) - \hat{\mathbf{B}}_n^i(t_1)| = |n^{-1}B_n^i(nt_2) - n^{-1}B_n^i(nt_2)| \leq |t_2 - t_1| \, .$$

Hence, $\{(\hat{\mathbf{B}}_n^1, \ldots, \hat{\mathbf{B}}_n^m)\}$ has a convergent subsequence $\{(\hat{\mathbf{B}}_{n_k}^1, \ldots, \hat{\mathbf{B}}_{n_k}^m)\}$ in $C([0,T], \mathbb{R}^k, U)^m$ for every T. Suppose that

$$(\hat{\mathbf{B}}_{n_k}^1, \ldots, \hat{\mathbf{B}}_{n_k}^m) \to (\hat{\mathbf{B}}^1, \ldots, \hat{\mathbf{B}}^m) \tag{2.26}$$

as $n_k \to \infty$ in (D^m, U).

We now use the assumed FCLT in (2.9) and this convergence to establish a FWLLN along the subsequence $\{n_k\}$. In particular, it follows that

$$\hat{\mathbf{X}}_{n_k} \Rightarrow 0\mathbf{e} \ ,$$

where $\hat{\mathbf{X}}_n \equiv (c_n/n)\mathbf{X}_n$ for $\mathbf{X}_n$ in (2.20). By the continuous-mapping approach,

$$\hat{\mathbf{Y}}_{n_k} \Rightarrow \psi_L(0\mathbf{e}) = 0\mathbf{e} \ ,$$

where $\hat{\mathbf{Y}}_n \equiv (c_n/n)\mathbf{Y}_n$ for $\mathbf{Y}_n$ in (2.22). Consequently, we must have $\hat{\mathbf{B}}^i = \mathbf{e}$ for all i for $\hat{\mathbf{B}}^i$ in (2.26). Since the same limit holds for all subsequences, we actually have the desired convergence in (2.25). We can then apply the continuous-mapping approach to establish that $\mathbf{X}_n \Rightarrow \mathbf{X}$ for $\mathbf{X}_n$ in (2.20), $\mathbf{X} = \mathbf{A} - \mathbf{N} + \eta\mathbf{e}$ and $\mathbf{N}$ in (2.14). ∎

In the general setting of Theorem 10.2.3, allowing limit processes with jumps, it remains to establish related stochastic-process limits for other queueing processes such as the workload and the waiting time.

Remark 10.2.3. *Bounds using other disciplines.* As shown by Wolff (1977), for the case in which all service times come from a single sequence of IID random variables, we can bound the queue length in the m-server FCFS model above stochastically by considering alternative service disciplines. For example, if the servers are allowed to have their own queues and arrivals are routed randomly or cyclically (in a round robin manner) to the servers, then the sum of the queue lengths stochastically dominates the queue length with the FCFS discipline. With the random or cyclic routing, we obtain heavy-traffic

FCLT's for the single servers under the assumptions of Theorem 10.2.1, with the same scaling as in Theorem 10.2.1. We apply Theorem 9.5.1 to treat the arrival processes to the separate servers. We note that, even though the scaling is the same, the limit processes are different. Thus the differences in the service disciplines can be seen in the heavy-traffic limit.

Following Loulou (1973), we can also bound the workload process in the m-server FCFS model below by the workload process in a single-server queue with the same total input and constant output rate m. Thus we can obtain a nondegenerate limiting lower bound for the scaled workload process using the scaling in Theorem 10.2.1. ∎

Remark 10.2.4. *Resource pooling with other disciplines.* From Theorem 10.2.3 and Remark 10.2.2, we see that, under minor regularity conditions, there is resource pooling in heavy traffic for the standard multi-server model with homogeneous servers and the FCFS discipline. As noted in Remark 10.2.3 above, there is *not* resource pooling in heavy traffic when each server has its own queue and random or cyclic routing is used. However, there is resource pooling in heavy traffic with the join-the-shortest-queue rule and many related service disciplines that pay only a little attention to the system state; see Foschini and Salz (1978), Reiman (1984b), Laws (1992), Kelly and Laws (1993), Turner (1996, 2000) Harrison and Lopez (1999) and Bell and Williams (2001). The heavy-traffic behavior of random routing with periodic load balancing is analyzed by Hjálmtýsson and Whitt (1998). Once again, the time scaling in the heavy-traffic limit provides useful insight: The time scaling shows how the reconfiguration or balancing times should grow with the traffic intensity ρ in order to achieve consistent performance. For more on the impact of heavy-tailed distributions on load balancing, see Harchol-Balter and Downey (1997). For more on the great gains from only a little choice, see Azar et al. (1994), Vvedenskaya et al. (1996), Mitzenmacher (1996), Turner (1998) and Mitzenmacher and Vöcking (1999). ∎

10.3. Infinitely Many Servers

It may happen that there is a very large number of servers. An extreme case of a large number of servers is the case of infinitely many servers. With infinitely many servers, heavy-traffic is achieved by letting the arrival rate approach infinity. With infinitely many servers, we assume that the system starts out empty and that the service times are IID and independent of the arrival process.

Remark 10.3.1. *The power of infinite-server approximations.* More generally, when the load in a multi-server queue is light or there are a very large number of servers, it may be helpful to consider infinite-server models as approximations. The infinite-server assumption is attractive because the infinite-server model is remarkably tractable. For infinite-server models, it is even possible to obtain useful expressions for performance measures with time-varying arrival rates; e.g., see Eick, Massey and Whitt (1993), Massey and Whitt (1993) and Nelson and Taaffe (2000).

Infinite-server models are also called *offered-load models*, because they describe the load (number of busy servers) that would results if there were no capacity constraints, so that no customers are delayed or lost. At first glance, offered-load models may seem unrealistic as direct system models, but they offer great potential for engineering because they are tractable and because they actually do not differ greatly from associated, more complicated, delay and loss models when the capacity and offered load are large. The idea is to engineer so that the probability that the offered load exceeds capacity is sufficiently small; see Jennings, Mandelbaum, Massey and Whitt (1996), Leung, Massey and Whitt (1994), Massey and Whitt (1993, 1994a, b), Duffield and Whitt (1997, 1998, 2000) and Duffield, Massey and Whitt (2001). ∎

10.3.1. Heavy-Traffic Limits

Following Glynn and Whitt (1991), we show how to apply the continuous-mapping approach to establish heavy-traffic stochastic process limits for infinite-server queues when the service-time distribution takes values in a finite set. As shown by Borovkov (1967), heavy-traffic limits can be established for general $G/GI/\infty$ queues with general service-time distributions, but the argument is more elementary when the service-time distributions take values in a finite set. For other heavy-traffic limits and approximations for infinite-server queues, see Fleming and Simon (1999) and Krichagina and Puhalskii (1997).

The key observation from Glynn and Whitt (1991) is that, when the system is initially empty, the queue length (again number in system) is simply related to the arrival process when the service time is deterministic. Let $\{A(t) : t \geq 0\}$ and $\{Q(t) : t \geq 0\}$ be the arrival and queue-length processes, respectively. When the service time is x for all customers, the queue length at time t is simply

$$Q(t) = A(t) - A(t-x) \ , \tag{3.1}$$

where we adopt the convention throughout this section that stochastic processes evaluated at negative arguments are identically 0; i.e., here $A(u) = 0$ for $u < 0$.

To exploit this simple representation more generally, we assume that all customers have service times in the finite set $\{x_1, \ldots, x_m\}$. We say that a customer with service time x_i is of type i and we let $A_i(t)$ count the number of type-i arrivals in the time interval $[0, t]$. We then let $D_i(t)$ count the number of type-i departures in $[0, t]$ and $Q_i(t)$ be the type-i queue length at time t. As in (3.1), we have the simple relations

$$\begin{aligned} D_i(t) &= A_i(t - x_i) \\ Q_i(t) &= A_i(t) - A_i(t - x_i), \quad t \geq 0 \ . \end{aligned}$$

We can also treat the workload process, representing the sum of the remaining service times of all customers in the system. To do so, we let $R_i(t, y)$ be the number of type-i customers with remaining service time greater than y in the system at time t and let $L_i(t)$ be the type-i workload at time t. Then we clearly have

$$R_i(t, y) = A_i(t) - A_i(t - [x_i - y]^+), \quad t \geq 0 \ ,$$

and

$$L_i(t) = \int_0^{x_i} R_i(t, y) \, dy \ .$$

We introduce the remaining-service-time processes as a means to treat the workload process, but the remaining-service-time process is of interest to describe the relevant state of the queue. To characterize the state of the customers in service, we need to consider the remaining-service-time $R_i(t, y)$ as a function of y for $y \geq 0$. In the $GI/GI/\infty$ model, when we append the elapsed interarrival time to the remaining-service-time process $\{R_i(t, y) : y \geq 0\}$, we obtain a Markov process as a function of t. In the heavy-traffic limit with a renewal arrival process, the interarrival time in process becomes negligible, so that the heavy-traffic limit for the remaining-service-time process is a Markov process as a function of t.

To apply the continuous mapping approach to treat the workload processes L_i, we need to regard $R_i(t, y)$ as a function mapping t into functions of y. We let the range be the subset of nonincreasing nonnegative functions in D with finite L_1 norm

$$\|x\|_1 \equiv \int_0^\infty |x(t)| \, dt \ . \tag{3.2}$$

Let $R_i : [0,\infty) \to (D, \|\cdot\|_1)$ be defined by $R_i(t) = \{R_i(t,y) : y \geq 0\}$. Since $(D, \|\cdot\|_1)$ is a Banach space, we can use the M_1 topology on the space $D([0,\infty), (D, \|\cdot\|_1))$, as noted in Section 11.5.

Let $A(t)$, $D(t)$, $Q(t)$, $R(t,y)$, $R(t)$ and $L(t)$ denote the associated m-dimensional vectors, e.g., $A(t) \equiv (A_1(t), \ldots, A_m(t))$, and let $\tilde{A}(t)$, etc. denote the associated partial sums, e.g., $\tilde{A}(t) \equiv A_1(t) + \cdots + A_m(t)$.

We form a sequence of infinite-server systems by scaling time in the original arrival process. Specifically, let

$$\begin{aligned}
A_{n,i}(t) &\equiv A_i(nt), \\
D_{n,i}(t) &\equiv A_{n,i}(t - x_i), \\
Q_{n,i}(t) &\equiv A_{n,i}(t) - A_{n,i}(t - x_i), \\
R_{n,i}(t,y) &\equiv A_{n,i}(t) - A_{n,i}(t - [x_i - y]^+), \\
L_{n,i}(t) &\equiv \int_0^{x_i} R_{n,i}(t,y)\, dy \ , \quad t \geq 0 \ .
\end{aligned} \tag{3.3}$$

Now we define associated random elements of D by letting

$$\begin{aligned}
\mathbf{A}_n^i(t) &\equiv c_n^{-1}[A_{n,i}(t) - \lambda_i nt], \\
\mathbf{D}_n^i(t) &\equiv c_n^{-1}[D_{n,i} - nd_i(t)], \\
\mathbf{Q}_n^i(t) &\equiv c_n^{-1}[Q_{n,i}(t) - nq_i(t))], \\
\mathbf{L}_n^i(t) &\equiv c_n^{-1}[L_{n,i}(t) - nl_i(t)], \quad t \geq 0 \ , \\
\mathbf{R}_n^i(t) &\equiv \{c_n^{-1}[Q_{n,i}(t,y) - nr_i(t,y)] : y \geq 0\} \ ,
\end{aligned} \tag{3.4}$$

where the translation functions are

$$\begin{aligned}
d_i(t) &\equiv \lambda_i([t - x_i]^+), \\
q_i(t) &\equiv \lambda_i(t - [t - x_i]^+), \\
l_i(t) &\equiv \int_0^{x_i} q_i(t,y)\, dy, \\
r_i(t,y) &\equiv \lambda_i(t - [t - [x_i - y]^+]^+) \ .
\end{aligned} \tag{3.5}$$

Let the associated vector-valued random elements of D^m and partial sums in D be

$$\begin{aligned}
\mathbf{A}_n &\equiv (\mathbf{A}_n^1, \ldots, \mathbf{A}_n^m), \\
\mathbf{D}_n &\equiv (\mathbf{D}_n^1, \ldots, \mathbf{D}_n^m), \\
\mathbf{Q}_n &\equiv (\mathbf{Q}_n^1, \ldots, \mathbf{Q}_n^m), \\
\mathbf{L}_n &\equiv (\mathbf{L}_n^1, \ldots, \mathbf{L}_n^m), \\
\mathbf{R}_n &\equiv (\mathbf{R}_n^1, \ldots, \mathbf{R}_n^m)
\end{aligned} \tag{3.6}$$

and

$$\begin{aligned}
\tilde{\mathbf{A}}_n &\equiv \mathbf{A}_n^1 + \cdots + \mathbf{A}_n^m, \\
\tilde{\mathbf{D}}_n &\equiv \mathbf{D}_n^1 + \cdots + \mathbf{D}_n^m, \\
\tilde{\mathbf{Q}}_n &\equiv \mathbf{Q}_n^1 + \cdots + \mathbf{Q}_n^m, \\
\tilde{\mathbf{L}}_n &\equiv \mathbf{L}_n^1 + \cdots + \mathbf{L}_n^m, \\
\tilde{\mathbf{R}}_n &\equiv \mathbf{R}_n^1 + \cdots + \mathbf{R}_n^m \ . \qquad (3.7)
\end{aligned}$$

We now can establish the basic heavy-traffic stochastic-process limit. For that purpose, let $\theta_s : D^m \to D^m$ denote the shift operator, defined by

$$\theta_s(x)(t) \equiv x(t+s), \quad t+s \geq 0 \ ,$$

with $\theta_s(x)(t) = 0$ for $t + s < 0$, all for $t \geq 0$. We exploit the fact that the shift operator θ_s is continuous for $s \leq 0$. However, note that the shift operator θ_s is *not* continuous for $s > 0$. To see this, let $s = 1$, $x = I_{[1,\infty)}$ and $x_n = I_{[1+n^{-1},\infty)}$.

Let the components of the vector-valued limit processes also be indexed by superscripts.

Theorem 10.3.1. (heavy-traffic limit for infinite-server queues) *If*

$$\mathbf{A}_n \Rightarrow \mathbf{A} \quad \textit{in} \quad (D^m, WM_1) \ ,$$

where

$$P(Disc(\mathbf{A}^i) \cap Disc(\theta_s(\mathbf{A}^j) = \phi) = 1 \qquad (3.8)$$

for all i, j and $s \leq 0$ for which $i \neq j$ or $i = j$ and $s \neq 0$, then

$$\begin{aligned}
&(\mathbf{A}_n, \tilde{\mathbf{A}}_n, \mathbf{D}_n, \tilde{\mathbf{D}}_n, \mathbf{Q}_n, \tilde{\mathbf{Q}}_n, \mathbf{L}_n, \tilde{\mathbf{L}}_n, \mathbf{R}_n, \tilde{\mathbf{R}}_n) \\
&\quad \Rightarrow (\mathbf{A}, \tilde{\mathbf{A}}, \mathbf{D}, \tilde{\mathbf{D}}, \mathbf{Q}, \tilde{\mathbf{Q}}, \mathbf{L}, \tilde{\mathbf{L}}, \mathbf{R}, \tilde{\mathbf{R}})
\end{aligned}$$

in $D([0,\infty), \mathbb{R})^{4(m+1)} \times D([0,\infty), \mathbb{R}, \|\cdot\|_1))^{m+1}$ with the WM_1 topology, where

$$\begin{aligned}
\mathbf{D}^i(t) &\equiv \mathbf{A}^i(t - x_i), \\
\mathbf{Q}^i(t) &\equiv \mathbf{A}^i(t) - \mathbf{A}^i(t - x_i), \\
\mathbf{L}^i(t) &\equiv \int_0^{x_i} \mathbf{R}^i(t)(y)\, dy, \\
\mathbf{R}^i(t)(y) &\equiv \mathbf{A}^i(t) - \mathbf{A}^i(t - [x_i - y]^+) \\
\tilde{\mathbf{A}}(t) &\equiv \mathbf{A}^1(t) + \cdots + \mathbf{A}^m(t), \\
\tilde{\mathbf{D}}(t) &\equiv \mathbf{D}^1(t) + \cdots + \mathbf{D}^m(t), \\
\tilde{\mathbf{Q}}(t) &\equiv \mathbf{Q}^1(t) + \cdots + \mathbf{Q}^m(t), \\
\tilde{\mathbf{L}}(t) &\equiv \mathbf{L}^1(t) + \cdots + \mathbf{L}^m(t), \\
\tilde{\mathbf{R}}(t) &\equiv \mathbf{R}^1(t) + \cdots + \mathbf{R}^m(t) \ . \qquad (3.9)
\end{aligned}$$

Proof. We apply the continuous mapping theorem, Theorem 3.4.3, with a succession of maps that are measurable and almost surely continuous with respect to the limit process, by virtue of condition (3.8). The linearity of the model means that the scaled processes are related by the same maps the are used to construct the original processes. The maps are shown applied to the limit process $\mathbf{A}$ in (3.9). Note in particular that the map taking A_i into R_i, and thus $\mathbf{A}_n^i$ into $\mathbf{R}_n^i$, is continuous. ∎

Remark 10.3.2. *Stationarity after finite time.* If the limit process $\mathbf{A}$ has stationary increments, then the process $\theta_s(\mathbf{X}) \equiv \{\mathbf{X}(t+s) : t \geq 0\}$ is a stationary process when $s \geq max\{x_1, \ldots, x_m\}$ and $\mathbf{X}$ is any of the following limit processes: $\mathbf{A}(t+u) - \mathbf{A}(u)$, $\mathbf{Q}(t)$, $\mathbf{D}(t+u) - \mathbf{D}(u)$, $\mathbf{R}(t)(y)$ and $\mathbf{L}(t)$ for $u \geq 0$ and $y > 0$. Hence these limit processes have the property that they reach steady state in finite time (as the original processes do with a Poisson arrival process). If, in addition, $\mathbf{A}$ is a Gaussian process, such as a fractional Brownian motion, then these processes are stationary Gaussian processes. ∎

10.3.2. Gaussian Approximations

We will now focus on the common case in which the conditions of Theorem 10.3.1 hold with $c_n = \sqrt{n}$ and $\mathbf{A}$ an m-dimensional Brownian motion. As noted in Remark 10.3.2 above, when $\mathbf{A}$ is an m-dimensional Brownian motion, the limit processes $\mathbf{D}$, $\mathbf{Q}$, $\mathbf{L}$ and $\mathbf{R}(\cdot)(y)$ are all Gaussian processes. Assuming appropriate uniform integrability in addition to the condition of Theorem 10.3.1, so that the variances converge (see p. 32 of Billingsley (1968)), we can relate the covariance matrix Σ_A of the Brownian motion $\mathbf{A}$ to the original arrival processes. In particular, under that regularity condition,

$$\Sigma_{A,i,j} = \lim_{t \to \infty} t^{-1} cov(A_i(t), A_j(t)) \ . \tag{3.10}$$

We now describe the relatively simple stationary Gaussian approximations that hold for the aggregate departure, queue-length and workload processes when the limit process $\mathbf{A}$ is an m-dimensional Brownian motion and $t \geq max\{x_1, \ldots, x_m\}$:

$$\tilde{D}(t+s) - \tilde{D}(t) \approx N(\sum_{i=1}^{m} \lambda_i s, \sum_{i=1}^{m} \sum_{j=1}^{m} \Sigma_{D,i,j}),$$

$$\begin{aligned}\tilde{Q}(t) &\approx N(\sum_{i=1}^{m}\lambda_i x_i, \sum_{i=1}^{m}\sum_{j=1}^{m}\Sigma_{Q,i,j}),\\ \tilde{L}(t) &\approx N(\sum_{i=1}^{m}\lambda_i x_i^2/2, \sum_{i=1}^{m}\sum_{j=1}^{m}\Sigma_{L,i,j}) \ , \end{aligned} \tag{3.11}$$

where

$$\begin{aligned}\Sigma_{D,i,j} &\equiv \Sigma_{A,i,j}([s-|x_i-x_j|]^+),\\ \Sigma_{Q,i,j} &\equiv \Sigma_{A,i,j}(x_i \wedge x_j)),\\ \Sigma_{L,i,j} &\equiv \Sigma_{A,i,j}[(x_i \vee x_j)(x_i \wedge x_j)/2 - (x_i \wedge x_j)^3/6] \ . \end{aligned} \tag{3.12}$$

To obtain the covariance terms $\Sigma_{L,i,j}$ for the workload, we use the representation

$$\Sigma_{L,i,j} = \int_0^{x_i}\int_0^{x_j} cov[\mathbf{A}^i(t) - \mathbf{A}^i(t-x_i+y), \mathbf{A}^j(t) - \mathbf{A}^j(t-x_j+z)]\, dz\, dy \ .$$

The assumption of Theorem 10.3.1 starting out with a limit for $\mathbf{A}_n \equiv (\mathbf{A}_n^1, \ldots, \mathbf{A}_n^m)$ is natural when there are m classes of jobs each with their characteristic deterministic service time. However, we are often interested in a single arrival process, with each successive arrival being randomly assigned a service time, which we here take to be from the finite set $\{x_1, \ldots, x_m\}$. Sufficient conditions for the condition in Theorem 10.3.1 in that setting follow from Theorem 9.5.1 on split streams. Suppose that the limit process $(\tilde{\mathbf{A}}, \mathbf{S})$ there is a centered $(m+1)$-dimensional Brownian motion with covariance matrix Σ. Then the limit process $\mathbf{A}$ here is a centered m-dimensional Brownian motion with covariance matrix

$$\Sigma_{A,i,j} = \lambda\Sigma_{i,j} + p_i\lambda^{1/2}\Sigma_{i,m+1} + p_j\lambda^{1/2}\Sigma_{j,m+1} + p_i p_j \Sigma_{m+1,m+1} \ . \tag{3.13}$$

An important application of the setting above occurs when the service times are IID with distribution $P(V = x_i) = p_i$, and independent of the arrival process. If we further assume that the limit process $\tilde{\mathbf{A}}$ in Theorem 9.5.1 is $\lambda^{3/2}\sigma B$, where $\mathbf{B}$ is standard Brownian motion, as occurs for a renewal process when the interrenewal times have mean λ^{-1} and variance σ^2, then the covariance terms become $\Sigma_{i,i} = p_i(1-p_i)$ for $i \le m$, $\Sigma_{m+1,m+1} = \lambda^3\sigma^2$, $\Sigma_{i,j} = -p_i p_j$ for $i, j \le m$ and $i \ne j$ and $\Sigma_{i,m+1} = 0$ for $i \le m$. Then the covariance function of the limiting Gaussian process $\mathbf{Q}$ in Theorem 10.3.1 can be represented as

$$\begin{aligned} cov(\mathbf{Q}(s), \mathbf{Q}(s+t)) &= \lambda\int_0^s H(u)H^c(t+u)\, du\\ &\quad + \sigma^2\lambda^3\int_0^s H^c(t+u)H^c(u)\, du \ , \end{aligned} \tag{3.14}$$

where $H(t) \equiv P(V \le t)$ is the service-time cdf and $H^c(t) \equiv 1 - H(t)$ is the associated ccdf.

Borovkov (1967) established a heavy-traffic limit justifying (3.14) for general service-time distributions. Borovkov assumes that the scaled arrival process converges to $\lambda^{3/2}\sigma\mathbf{B}$, where $\mathbf{B}$ is standard Brownian motion. For general service-time cdf H and Brownian arrival-process limit, the heavy-traffic limit shows that the stationary queue length in the $G/GI/\infty$ model is approximately distributed according to

$$Q(\infty) \approx N(\gamma, \gamma z_Q) \ , \tag{3.15}$$

where $\gamma \equiv \lambda/\mu$ is the *total offered load* and

$$\begin{aligned} z_Q &= \mu \int_0^\infty H(u)H^c(u)\,du + \lambda^2\mu\sigma^2 \int_0^\infty H^c(u)^2\,du \\ &= 1 + (c_U^2 - 1)\mu \int_0^\infty H^c(u)^2\,du \ , \end{aligned} \tag{3.16}$$

with $c_U^2 \equiv \lambda^2\sigma^2$.

To put the normal approximation in (3.15) in perspective, the steady-state mean is exactly $EQ(\infty) = \gamma$ by Little's law, $L = \lambda W$. With a Poisson arrival process, $Q(\infty)$ has a Poisson distribution with mean γ, which is asymptotically normal as λ goes to infinity with fixed service rate μ. With a Poisson arrival process, $c_U^2 = c_u^2 = 1$; when $c_U^2 = 1$, $z_Q = 1$ and $\sigma_Q^2 = \gamma$.

More generally, we see that the asymptotic variance is γz_Q, so that $Q(\infty)$ tends to differ from the mean γ by amounts of order $O(\sqrt{\gamma z_Q})$ as γ gets large. The variance scale factor z_Q is called the *asymptotic peakedness*. More generally, the *peakedness* is the ratio

$$Var(Q(\infty)/EQ(\infty) = Var(Q(\infty))/\gamma \ .$$

The peakedness and the asymptotic peakedness are often used in approximations of loss and delay systems with finitely many servers; see Eckberg (1983, 1985), Whitt (1984a, 1992b) and Srikant and Whitt (1996).

A key quantity in the asymptotic peakedness is the integral $\mu \int_0^\infty H^c(u)^2\,du$. It varies from 1 when H is the cdf of the unit point mass on μ^{-1} to 0. That integral tends to *decrease* as the service-time distribution gets more variable with the mean held fixed. Note that the effect of sevice-time variability on the asymptotic peakedness z_Q depends on the sign of $(c_U^2 - 1)$.

Remark 10.3.3. *Exponential service times.* When the space scaling is by $c_n = \sqrt{n}$, the limit process for the arrival process is Brownian motion and

the service times are exponential with ccdf $H^c(t) = e^{-\mu t}$, the limiting Gaussian process $\mathbf{Q}$ becomes an Ornstein-Uhlenbeck diffusion process with infinitesimal mean $-\mu x$ and diffusion coefficient $\gamma\mu(c_u^2+1)$ for $\gamma = \sum_{i=1}^m \lambda_i x_i$. Direct heavy-traffic limits for the $M/M/\infty$ and $GI/M/\infty$ models were established by Iglehart (1965) and Whitt (1982c). When H is exponential, $\mu \int_0^\infty H^c(u)\,du = 1/2$, so that $z_Q = (c_U^2+1)/2$. ∎

Remark 10.3.4. *Prediction with non-exponential service times.* The fact that the limit process is a Markov process when the limit process for the arrival process is Brownian motion and the service times are exponential implies that, for exponential service times, in the heavy-traffic limit the future evolution of the process depends on the past only through the present state. In the heavy-traffic limit there is no benefit from incorporating additional information about the past.

However, that is not the case for non-Lévy limit processes for the arrival process and non-exponential service-time distributions. Then the elapsed service times of customers in service give information about the remaining service times of these customers.

As can be seen from the covariance formula in (3.14), the transient behavior depends on the full service-time cdf.

Example 10.3.1. *Transient behavior in the $M/GI/\infty$ queue.* Suppose that at some instant in steady state there happen to be n busy servers in an $M/GI/\infty$ model. In the case of a Poisson arrival process, we can decribe the future transient behavior, because, conditional on $Q(0) = n$, the n elapsed service times and the n residual service times are each distributed as n IID random variables with the stationary-excess cdf associated with the service-time cdf H, i.e.,

$$H_e(t) = \mu \int_0^t H^c(u)\,du, \quad t \geq 0\ . \tag{3.17}$$

Moreover, the number of new arrivals after time 0 that are still in service at a later time t is independent of the customers initially in service and has a Poisson distribution with mean $\lambda\mu^{-1}H_e(t)$; see Duffield and Whitt (1997).

As a consequence, it is elementary to compute the mean and variance of the conditional number of customers in the system at any future time, given the initial number n. If the arrival rate and offered load are large, then the number of customers in the system is likely to be large. From the properties above, we obtain a normal approximation refining the conditional mean. We can exploit this structure to study various control schemes to recover from rare congestion events. ∎

10.4. An Increasing Number of Servers

We now want to consider a third heavy-traffic limiting regime for multi-server queues. Just as we treated queues with superpositions of an increasing number of arrival processes in Section 9.8, we now want to treat queues with an increasing number of servers. Now we let the number m of servers go to infinity as the traffic intensity ρ approaches 1, the critical value for stability.

The m-server model we consider now is the standard m-server queue with unlimited waiting room and the FCFS service discipline. Customers are assigned in order of arrival to the first available server, with some unspecified procedure to break ties. the service times are independent of the arrival process and come from a sequence of IID random variables with mean μ^{-1}. The arrival rate is λ and the traffic intensity is $\rho \equiv \lambda/\mu m$. Heavy traffic will be obtained by increasing the arrival rate, using simple time scaling, just as done for the infinite-server model. But now we increase m as well as λ.

Just as in Remark 9.8.2, when the number of servers becomes large, the relevant time scale for an individual customer becomes different from the time scale of the system. The times between successive departures from the queue become much shorter than the individual service times. If there are many input streams as well as many servers, the time scale for individual customers is consistently much longer than the time scale for the queue. Thus, as in Section 9.8, the short-time behavior of individual customers will affect the large-time behavior of the queue.

We can use the infinite-server model to determine what the appropriate limiting regime for the m-server model should be. We will apply the infinite-server model to show that, when there is space scaling by $c_n = n^H$ for $0 < H < 1$, we should have $m \to \infty$ and $\rho \to 1$ with

$$m^{1-H}(1-\rho) \to \beta \tag{4.1}$$

for $0 < \beta < \infty$. (As before, the common case is $H = 1/2$.) Note that the growth rate here is the same as for superposition arrival processes in (8.10).

10.4.1. Infinite-Server Approximations

More generally, we can use an infinite-server model to estimate how many servers are needed in a finite-server system in order to achieve desired quality of service. To do so, we let the infinite-server model have the same arrival and service processes as the m-server model. We can use the probability

that m or more servers are busy in the infinite-server model as a rough approximation for the same quantity in an m-server model.

For that purpose, we let $m = m_p$, where

$$P(Q(\infty) \geq m_p) = p \tag{4.2}$$

for a target tail probability p. We can use Theorem 10.3.1 to generate an approximation for the distribution of $Q(\infty)$ and determine how the threshold m_p should grow as the arrival rate increases. We assume that Theorem 10.3.1 holds with $c_n = n^H$ for $0 \leq H \leq 1$. Then

$$Q_n(t) \approx nq(t) + n^H \tilde{\mathbf{Q}}(t) \ . \tag{4.3}$$

If we focus on the steady-state behavior, which occurs for $t \geq max\{x_1, \ldots, x_m\}$ when the limit process $\mathbf{A}$ has stationary increments, then

$$Q_n(\infty) \approx n \sum_{i=1}^{m} \lambda_i x_i + n^H \tilde{\mathbf{Q}}(\infty) \ . \tag{4.4}$$

Let γ denote the total offered load, which here is

$$\gamma = EQ_n(\infty) = n \sum_{i=1}^{m} \lambda_i x_i \ . \tag{4.5}$$

Since the total offered load can be expressed in terms of the traffic intensity ρ and the number of servers m by $\gamma = m\rho$, we can replace (4.4) by

$$Q(\infty) \approx \rho m + (c\rho m)^H \tilde{\mathbf{Q}}(\infty) \ . \tag{4.6}$$

for a constant c. (For $\gamma = n \sum_{i=1}^{m} \lambda_i x_i$, $c = n$.)

Combining equations (4.2) and (4.6), we obtain the approximation

$$P(Q(\infty) \geq m) \approx P(\tilde{\mathbf{Q}}(\infty) \geq m(1-\rho)/(c\rho m)^H) \ . \tag{4.7}$$

Let x_p be the $(1-p)^{\text{th}}$ quantile of the distribution of $\tilde{\mathbf{Q}}(\infty)$, i.e.,

$$P(\tilde{\mathbf{Q}}(\infty) \geq x_p) = p \ . \tag{4.8}$$

Combining (4.7) and (4.8), we obtain $m_p(1-\rho) = (c\rho m_p)^H x_p$ or

$$m_p \approx (x_p (c\rho)^H / (1-\rho))^{1/(1-H)} \ . \tag{4.9}$$

Approximation (4.9) specifies the required number of servers, using the infinite-server constraint (4.2) and the heavy-traffic limit established in Theorem 10.3.1.

In the common case in which $c_n = n^{1/2}$ and $Q(\infty)$ has the normal approximation in (3.15), we have the approximation

$$P(Q(\infty) \geq m) \approx P(N(\gamma, \gamma z_Q) \geq m) = \Phi^c((m-\gamma)/\sqrt{\gamma z_Q}) \ , \qquad (4.10)$$

where $\gamma \equiv \rho m$ is again the offered load, z_Q is the asymptotic peakedness and Φ^c is the standard normal ccdf. Then we obtain the special case of (4.9)

$$m_p \approx ((x_p \sqrt{\rho z_Q})/(1-\rho))^2 \ . \qquad (4.11)$$

Thus, we have developed an approximation for the required number of servers in an m-server system based on an infinite-server-model approximation. This same approach applies to multi-server queues with time-varying arrival rates; see Jennings, Mandelbaum, Massey and Whitt (1996).

Note that equations (4.9) and (4.11) show that there is increased service efficiency as the number m of servers increases. The traffic intensity at which the system can satisfy the performance constraint (4.2) increases as m increases; see Smith and Whitt (1981) and Whitt (1992) for further discussion.

From (4.9) and (4.11), we also can determine the rate at which m should grow as $\rho \to 1$ in m-server systems so that the probability of delay approaches a nondegenerate limit (a limit p with $0 < p < 1$). In the infinite server model, we meet the tail probability constraint (4.2) as the arrival rate increases (by simple time scaling as in the previous subsection) if $m \to \infty$ and $\rho \equiv \gamma/m \to 1$ with m and ρ related by (4.9). In other words, we obtain (4.1) as an estimate of the way in which m should be related to ρ as $m \to \infty$ and $\rho \to 1$ in the m-server model.

10.4.2. Heavy-Traffic Limits for Delay Models

Of course, the infinite-server model is just an approximation. It remains to establish heavy-traffic limits as $\rho \to 1$ and $m \to \infty$ with (4.1) holding in an m-server model. However, this third limiting regime is more complicated, evidently requiring methods beyond the continuous-mapping approach. We will briefly summarize heavy-traffic limits established in this regime by Halfin and Whitt (1981) for the $GI/M/m$ model, having exponential service times and $H = 1/2$. Puhalskii and Reiman (2000) established heavy-traffic limits, with more complicated limit processes, for the more general $GI/PH/m$ model with phase-type service-time distributions.

Let $Q_m(t)$ be the queue length at time t and let $Q_m(\infty)$ be the steady-state queue length in a standard $GI/M/m$ model with m servers. Let the

interarrival times be constructed from a sequence $\{U_k : k \geq 1\}$ of IID random variables with mean 1 and SCV c_u^2. When the number of servers is m, let the interarrival times be

$$U_{m,k} \equiv U_k/\lambda_m \ , \tag{4.12}$$

so that the arrival rate in model m is λ_m. Let the individual service rate be μ for all m, so that the traffic intensity as a function of m is $\rho_m = \lambda_m/\mu m$.

We now state the two main results from Halfin and Whitt (1981) without proof. The first concerns the steady-state distribution. The second is a FCLT.

Theorem 10.4.1. (necessary and sufficient conditions for asymptotically nondegenerate delay probability) *For the family of $GI/M/m$ models specified above,*

$$\lim_{m\to\infty} P(Q_m(\infty) \geq m) = p, \quad 0 < p < 1 \ , \tag{4.13}$$

if and only if the arrival rate λ_m increases with m so that

$$\lim_{m\to\infty} (1 - \rho_m)m^{1/2} = \beta, \quad 0 < \beta < \infty \ , \tag{4.14}$$

in which case

$$p = [1 + \xi\sqrt{2\pi}\Phi(\xi)exp(\xi^2/2)]^{-1} \ , \tag{4.15}$$

where

$$\xi = 2\beta/(1 + c_u^2) \ . \tag{4.16}$$

Moreover, if (4.14) *holds, then*

$$m^{1/2}(Q_m(\infty) - m) \Rightarrow Z \quad in \quad \mathbb{R} \ , \tag{4.17}$$

where

$$P(Z \geq 0) = p, \quad P(Z > x|Z \geq 0) = e^{-x\xi} \tag{4.18}$$

and

$$P(Z \leq x|Z \leq 0) = \Phi(x + \xi)/\Phi(\xi) \tag{4.19}$$

for ξ in (4.16).

To state the FCLT, we construct random elements of $D \equiv D([0,\infty), \mathbb{R})$ by letting

$$\mathbf{Q}_m(t) \equiv m^{-1/2}(Q_m(t) - m), \quad t \geq 0 \ . \tag{4.20}$$

There is no time scaling in (4.20) because the arrival rate λ_m is allowed to grow directly.

Theorem 10.4.2. (Heavy-traffic FCLT with an increasing number of servers) *If* (4.14) *holds and* $\mathbf{Q}_m(0) \Rightarrow \mathbf{Q}(0)$ *in* $\mathbb{R}$, *then*

$$\mathbf{Q}_m \Rightarrow \mathbf{Q} \quad in \quad (D, J_1) \quad as \quad m \to \infty \ , \tag{4.21}$$

where $\mathbf{Q}$ *is a diffusion process starting at* $\mathbf{Q}(0)$ *with infinitesimal mean*

$$m(x) = \begin{cases} -\mu\beta, & x \geq 0 \\ -\mu(x+\beta), & x < 0 \ , \end{cases}$$

diffusion coefficient

$$\sigma^2(x) = \mu(1 + c_u^2) \ .$$

and the steady-state distribution of Z *in Theorem* 10.4.1.

As indicated earlier, generalizations of Theorem 10.4.2 to $GI/PH/m$ queues have been established by Puhalskii and Reiman (2000). Generalizations to Markovian service networks with time-varying arrivals have been established by Mandelbaum, Massey and Reiman (1998). Approximations for the steady-state queue-length distribution and other steady-state distributions in $GI/GI/m$ models based partly on Theorem 10.4.1 are discussed in Whitt (1992, 1993a).

A large number of servers also affects the departure process from a queue. As shown by Whitt (1984f), under regularity conditions, even with general service-time distributions the departure process can be approximated by a Poisson process. As with the superposition arrival process, the Poisson property applies in a relatively short time scale.

Remark 10.4.1. *State dependence.* Even when restricting attention to simple Markovian $M/M/m$ queues, the limit process in Theorem 10.4.2 differs significantly from the limit processes when $m = 1$ and $m = \infty$. As shown in Section 10.2, the heavy-traffic limit for any fixed m is the same as for $m = 1$, but if we let $m \to \infty$ so that (4.1) holds, then we obtain a limiting diffusion process with a *nonlinear drift function*, which shows that there is significant state-dependence.

In contrast, the limit processes for single-server and infinite-server queues have essentially linear drift. For the single-server queues, there is constant drift, modified only by the reflection at the barriers. As a consequence of the scaling, the barrier disappears in the heavy-traffic limit for infinite-server models. Then the limit process has linear drift.

State dependent behavior may be important to capture in queueing models. In addition to the state-dependent consequence of multiple servers,

state-dependence occurs when there is balking (customers refusing to join a queue when it is congested) or reneging (customers abandoning the queue after waiting a long time). See Garnett, Mandelbaum and Reiman (2000) Ward and Glynn (2001) for heavy-traffic limits for queues with reneging.

It is natural to model state-dependent behavior directly by birth-and-death processes. We can then establish heavy-traffic limits in which the birth-and-death processes converge to diffusion processes. However, diffusion processes are continuous analogs of birth-and-death processes, so we may not gain much from such a limit, if we cannot show that the same limit holds for more general processes. However, state-dependence is sufficiently complicated that we may have to be content doing all the analysis in a Markovian framework.

For discussions of heavy-traffic limits for Markovian queues with state-dependence, see Browne and Whitt (1995), Pats (1994) and Mandelbaum and Pats (1995, 1998).

State dependence also arises when the service times and arrival times depend on the level of congestion; see Whitt (1990) for heavy-traffic limits in that setting.

10.4.3. Heavy-Traffic Limits for Loss Models

There is a heavy-traffic story for m-server loss models that closely parallels the heavy-traffic limits for m-server delay models just considered. We will consider $G/M/m$ loss models, in which all arrivals finding the m servers busy are blocked and lost, without affecting future arrivals. We will assume that the scaled arrival process satisfies a FCLT. Let A be a rate-1 counting process and let

$$\mathbf{A}_\lambda(t) \equiv \lambda^{-1/2}(A(\lambda t) - \lambda t), \quad t \geq 0 \ . \tag{4.22}$$

The first heavy-traffic limit is for the blocking probability. Let B_m be the blocking probability (the long-run proportion of arrivals that are blocked) as a function of the number of servers, m. The first heavy-traffic theorem is a *local limit theorem* due to Borovkov (1976). Related work is discussed in Whitt (1984a). A simple proof for the classical $M/M/m$ Erlang loss model is given in the appendix there. More detailed asymptotics for the Erlang model is contained in Jagerman (1974).

Theorem 10.4.3. (heavy-traffic limit for the blocking probability) *Consider a sequence of $GI/M/m$ loss models indexed by m with fixed individual service rate μ and arrival process $\{A(\lambda t) : t \geq 0\}$. Suppose that $\mathbf{A}_\lambda \Rightarrow \sigma_A \mathbf{B}$*

as $\lambda \to \infty$ *for* $\mathbf{A}_\lambda$ *in* (4.22) *and* $\mathbf{B}$ *standard Brownian motion, which for the renewal arrival process here is equivalent to the interarrival time having finite variance* σ_A^2. *If* $\lambda \to \infty$ *and* $m \to \infty$ *so that* (4.14) *holds or, equivalently, so that*

$$(m - \gamma)/\gamma \to \beta \tag{4.23}$$

for $\gamma \equiv \lambda/\mu$, *then*

$$\lim_{m \to \infty} \sqrt{\gamma} B_m = \sqrt{z}\, \phi(\beta/\sqrt{z})/\Phi(-\beta/\sqrt{z}) \ ,$$

where $z = (\sigma_A^2 + 1)/2$ *is the asymptotic peakedness in* (3.16).

There are significant differences between Theorems 10.4.3 and 10.4.1. Under the same conditions, the probability of delay in the $GI/M/m$ delay model approaches a nondegenerate limit, while the blocking probability in the $GI/M/m$ loss model is order $1/\sqrt{m}$ as $m \to \infty$. However, in both cases, the heavy-traffic limits show that the normal or critical operating regime when the offered load γ is large is $m = \gamma + c\sqrt{\gamma}$ for some constant c. The critical-loading regime is important for establishing asymptotic results for more general loss systems; e.g., see Hunt and Kelly (1989) and Reiman (1989, 1990b).

We now turn to the FCLT, in which we allow a general stationary arrival process. As above, we start with a rate-1 process. The FCLT is from Srikant and Whitt (1996), but it is closely related to Theorem 2 on p. 177 of Borovkov (1984).

Theorem 10.4.4. (FCLT for $G/M/m$ loss models) *Consider a sequence of* $G/M/m$ *loss models indexed by the number of servers,* m, *where the individual service rate is fixed at* μ. *Suppose that the arrival process is* $\{A(\lambda t) : t \geq 0\}$, *where* A *is a general stationary rate-*1 *process satisfying* $\mathbf{A}_\lambda \Rightarrow \sigma_A \mathbf{B}$ *as* $\lambda \to \infty$ *for* $\mathbf{A}_\lambda$ *in* (4.22) *and* $\mathbf{B}$ *standard Brownian motion. Suppose that* $m \to \infty$ *and* $\gamma \equiv \lambda/\mu \to \infty$ *with* (4.23) *holding and* $\gamma^{-1/2}(Q_m(0) - m) \Rightarrow y$, *where* $\{Q_m(t) : t \geq 0\}$ *is the queue-length process in model* m. *Then*

$$\mathbf{Q}_m \Rightarrow \mathbf{Q} \quad in \quad (D, J_1) \ ,$$

where

$$\mathbf{Q}_m(t) \equiv \gamma^{-1/2}(Q_m(t) - m), \quad t \geq 0 \ ,$$

and $\mathbf{Q}$ *is a reflected Ornstein-Uhlenbeck process with infinitesimal mean* $m(x) = -\mu(x + \beta)$ *for* $x \leq 0$, *infinitesimal variance* $\sigma^2(x) = \mu(1 + \sigma_A^2)$, *initial position* $\mathbf{Q}(0) = y$ *and instantaneous reflecting barrier above at* 0.

Remark 10.4.2. *Exponential service times.* In the case of exponential service times, the heavy-traffic limits for the infinite-server, delay and loss models involve essentially the same Ornstein-Uhlenbeck (OU) diffusion process. Of course, for the delay model, the diffusion acts like the OU diffusion only on part of its state space, i.e., when the servers are not all busy. The diffusion coefficient in the case of the infinite-server model in Remark 10.3.3 appears different only because the space scaling there is by $\sqrt{n}$ instead of by the square root of the offered load; the offered load there is $n\sum_{i=1}^{m}\lambda_i x_i$. ■

10.4.4. Planning Simulations of Loss Models

Just as in Section 5.8, the heavy-traffic stochastic-process limits here can be used to help plan simulations. Assuming that our goal is to estimate the long-run blocking probability, we can use the heavy-traffic limits to produce approximations for the blocking probability and the asymptotic variance of the estimator appearing in the formulas for the required simulation run length to achieve desired statistical precision. Estimators of the blocking probability were investigated by Srikant and Whitt (1996, 1999).

The *natural estimator* for the steady-state blocking probability B based on observations of the system over the time interval $[0,t]$ is

$$\hat{B}_N(t) \equiv L(t)/A(t) \ , \tag{4.24}$$

where $L(t)$ is the number of lost arrivals and $A(t)$ is the number of arrivals in $[0,t]$. Since we may know the arrival rate in a stationary arrival process with rate λ, an alternative *simple estimator* is

$$\hat{B}_S(t) \equiv L(t)/\lambda t \ . \tag{4.25}$$

Another alternative estimator can be based on Little's law, $L = \lambda W$. From Little's law, we have the relation

$$EQ(\infty) = \lambda(1-B)/\mu \ ; \tag{4.26}$$

here the effective arrival rate for admitted customers is $\lambda(1-B)$. Hence an alternative *indirect estimator* is

$$\hat{B}_I(t) \equiv 1 - \hat{m}(t)/\gamma \ , \tag{4.27}$$

where $\gamma \equiv \lambda/\mu$ is the offered load and $\hat{m}(t)$ is an estimator of the steady-state mean $EQ(\infty)$. (Since we are considering a simulation, we assume that

λ and μ are known.) It is natural for $\hat{m}(t)$ to be the sample mean

$$\hat{m}(t) \equiv t^{-1} \int_0^t Q(u)\, du \ , \tag{4.28}$$

As in Section 5.8, the required simulation run length t is proportional to the asymptotic variance of the estimator, denoted again by σ^2. (Let us use the criterion of absolute error.) However, the computational effort to simulate for time t is approximately proportional to λt, because that is the expected number of arrivals in the interval $[0, t]$. Hence it is natural to focus on the *workload factor* $\lambda\sigma^2$.

Srikant and Whitt (1996) apply heavy-traffic limits to develop approximations for the workload factors associated with the different estimators of the blocking probability. As a function of the basic parameters, they obtain the following approximation for the workload factor of the indirect estimator:

$$w_I \equiv w_I(m, \beta, c_a^2, c_s^2, z) \approx \frac{(c_a^2 + c_s^2)}{2} \psi_I(\beta/\sqrt{z}) \ , \tag{4.29}$$

where

$$\psi_I(x) \equiv w_I(\infty, x, 1, 1, 1) \tag{4.30}$$

is the *canonical workload factor* associated with the $M/M/m$ loss model asymptotically as $m \to \infty$, $\beta \equiv (\gamma - m)/\sqrt{\gamma}$ is the *scaled offered load* and z is the peakedness. Heavy-traffic enters in through our focus on the scaled offered load: the parameter β shows how the number of servers m is related to the offered load γ in the scaling associated with the FCLT for the cumulative-input process. The heavy-traffic limits suggest that the workload factor should depend upon the parameters γ and m primarily through the parameter β, provided that m is not too small and β is not too far from 0.

The approximations for the workload factors of the natural and simple estimators have the same form in (4.29) except the canonical workload factors are different. However, $\psi_S \approx \psi_N$, so that we henceforth restrict attention to the simple estimator and the indirect estimator.

Srikant and Whitt (1996) present theoretical arguments supporting the approximations, some of which involve heavy-traffic stochastic-process limits. They also present numerical comparisons supporting the approximations for $GI/GI/m$ models.

The approximations for the workload factors show remarkable statistical regularity. To demonstrate that there is indeed such statistical regularity, we plot the exact workload factors as functions of the number of servers

in the $M/M/m$ loss model. Since $c_a^2 = c_s^2 = z = 1$ in the $M/M/m$ loss model, only the two parameters m and β remain. In Figures 10.1 and 10.2 we plot the exact workload factors for the simple and indirect estimators as functions of β for several different values of m, ranging from $m = 25$ to $m = 800$.

The fact that the curves tend to fall on top of each other in the figures (except in one tail) shows that there is indeed remarkable statistical regularity. The different shape shows that the simple estimator is much more efficient in light loading, while the indirect estimator is much more efficient in heavy loading. Srikant and Whitt (1999) show that an empirically determined convex combination of these two estimators has the desirable qualities of both estimators, and is even more efficient.

Paralleling the heavy-traffic analysis for the single-server queue in Section 5.8, it is natural to ask how the required simulation run length changes as the number of servers, m, increases with the blocking probability held fixed. Interestingly, the story here is very different from the single-server queue. If we fix the blocking probability, then the required computational effort starting in steady state actually decreases to 0 as $m \to \infty$. On the other hand, the required run length to eliminate the initialization bias, starting empty, is approximately independent of system size. Hence, when the system size grows, a greater proportion of the computational effort must be devoted to eliminating the initialization bias. Alternatively, different initial conditions must be used in order to reduce initialization bias.

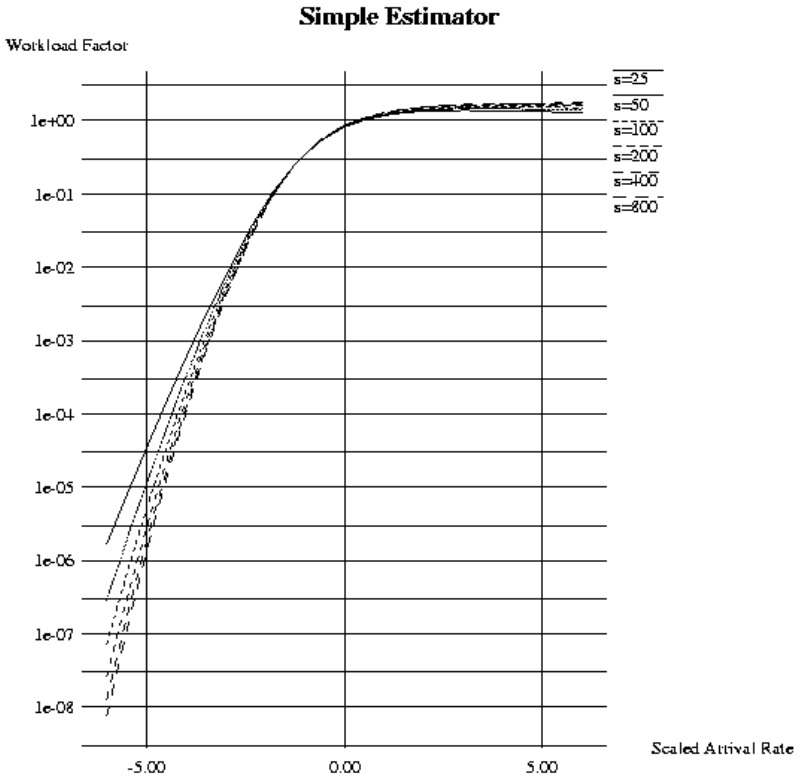

Figure 10.1: Workload factors $w_I \equiv \lambda\sigma_S^2$ for the simple estimator $\hat{B}_S(t)$ in the $M/M/m$ loss model with $\mu = 1$ as a function of the scaled arrival rate β for several values of m.

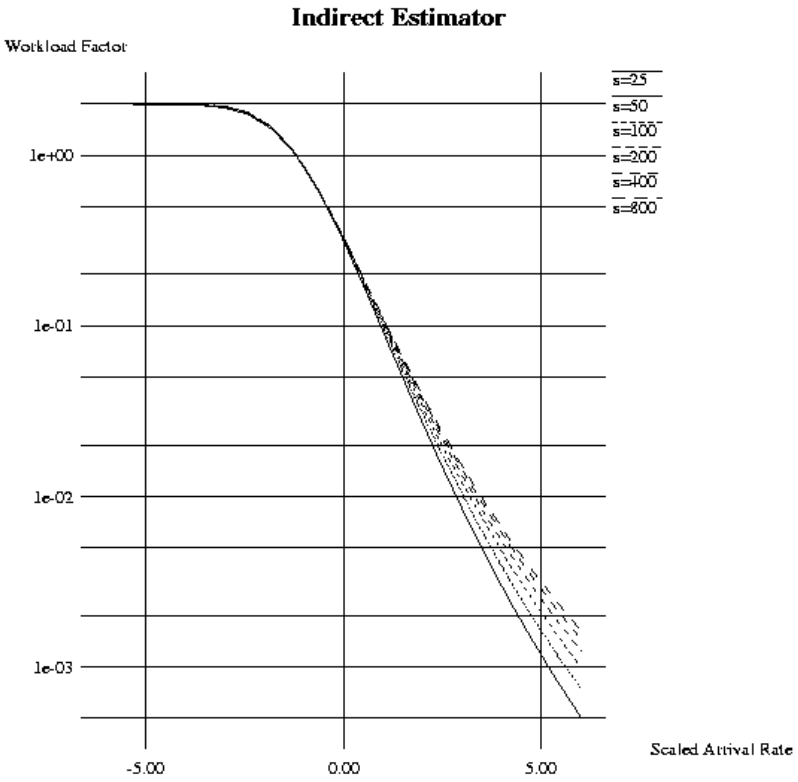

Figure 10.2: Workload factors $w_I \equiv \lambda\sigma_I^2$ for the as a function of the scaled arrival rate β for several values of m.