Chapter 5

Heavy-Traffic Limits for Queues

5.1. Introduction

In this chapter we include additional material on heavy-traffic limits for queues. The first two sections below supplement Chapter 8 in the book; the final section supplements Chapter 9 in the book.

In particular, Section 5.2 discusses general Lévy approximations for queues, obtained by considering a sequence of queueing models, exploiting the FCLT in Section 2.4 above and the continuous-mapping approach. Then Section 5.3 provides the missing proof to Theorem 8.3.1 in the book. Finally, Section 5.4, drawing upon Puhalskii (1994), shows how heavy-traffic limits for arrival, queue-length and departure processes can be used to establish associated limits for waiting-time and workload processes in single-server queues.

5.2. General Lévy Approximations

The Brownian and stable-Lévy approximations for queues in Chapters 5 and 8 in the book are robust approximations: The same approximation, characterized by only a few parameters, serves as an approximation for a large class of queueing models. We obtain the Brownian (stable-Lévy) approximation with light-tailed (heavy-tailed) distributions.

We can obtain a larger, more flexible, class of approximating processes if we consider stochastic-process limits based on a sequence of queueing models, where the input processes are allowed to change with the sequence index. Of course, we also can obtain the previous limit processes in this more general framework, but we can obtain new limit processes as well, which may be useful for applications.

Closely paralleling the previous sections, we can apply the continuous-mapping approach with the reflection map and a Lévy-process FCLT for double sequences in Theorem 2.4.1 here to obtain a stochastic-process limit for workload processes associated with a sequence of queueing models. When we allow the input processes to change in the limit, we can obtain stochastic-process limits without requiring heavy traffic.

As noted in Section 2.4, we obtain a large class of limit processes from the stochastic-process limits for partial sums from double sequences of random variables, with the variables in each sequence being IID. Indeed, the limit process for the net inputs can be an arbitrary Lévy process $\{L(t) : t \geq 0\}$. Of course, in applications it remains to determine the appropriate Lévy process. Since the Lévy process has stationary and independent increments, it suffices to specify the distribution of the random variable $L(1)$, which must be infinitely divisible. From (4.3) in Section 2.4, it suffices to specify the triple (b, σ^2, μ), where b is the centering constant, σ^2 is the Gaussian coefficient and μ is the Lévy measure. These limiting characteristics can be specified in approximations by exploiting the asymptotic relations in equations (4.10) – (4.12) in Section 2.4.

For applications, it is significant that there is a large class of reflected Lévy processes that are remarkably tractable. In particular, *a reflected Lévy process, constructed from a one-sided reflection, is tractable if the associated Lévy process has no negative jumps.* For example, the steady-state distribution can be characterized by its Laplace transform, which is often called the *generalized Pollaczek-Khintchine transform*, because the Pollaczek-Khintchine transform of the steady-state distribution of the workload process in the M/G/1 queue is a special case.

The original charachterization of the steady-state distribution of a reflected Lévy process for the case with no negative jumps is due to Zolotarev (1964); also see Section 24 of Takács (1967), Bingham (1975) and Kella and Whitt (1992c), especially Section 4(a). The short martingale proof in Kella and Whitt (1992c) is convenient.

When a Lévy process L has no negative jumps, the Lévy measure μ concentrates on $(0, \infty)$ and the bilateral Laplace-Stieltjes transform of $L(1)$

is well defined, with *Laplace exponent*

$$\begin{aligned}\psi(s) &\equiv \log Ee^{-sL(1)} \\ &= -bs + \frac{\sigma^2 s^2}{2} + \int_0^\infty (\exp(-sx) - 1 + sh(x))\mu(dx) \ . \end{aligned} \tag{2.1}$$

An important special case is a subordinator (totally skewed stable Lévy motion with $\beta = 1$ plus a negative drift, which is just (2.1) without the second Brownian term. Storage models with such Lévy net-input processes are analyzed directly in Chapter 4 of Prabhu (1998). With (2.1), we can conveniently characterize the Laplace transform of the steady-state distribution. The following is a generalization of Theorems 5.8.2 and 8.5.2 in the book.

Theorem 5.2.1. (generalized Pollaczek-Khintchine transform) *Let* $\{\phi_K(L)(t) : t \geq 0\}$ *be a reflected Lévy process, where* ϕ_K *is the two-sided reflection map,* $EL(1) < 0$, L *has no negative jumps and* L *has Laplace exponent* ψ *in (2.1).*

(a) If $K = \infty$, *then*

$$\lim_{t\to\infty} P(\phi_K(L))(t) \leq x) = H(x) \ , \tag{2.2}$$

where H *is a proper cdf with Laplace-Stieltjes transform*

$$\hat{h}(s) \equiv \int_0^\infty e^{-sx} dH(x) = \frac{s\psi'(0)}{\psi(s)} \ , \tag{2.3}$$

and ψ *is the Laplace exponent in (2.1).*

(b) If $K < \infty$, *then*

$$\lim_{t\to\infty} P(\phi_K(L))(t) \leq x) = \frac{H(x)}{H(K)}, \quad 0 \leq x \leq K \ , \tag{2.4}$$

for H *in (2.2).*

Example 5.2.1. *The special case of the M/G/1 queue.* The workload in unfinished service time in the M/G/1 queue is a reflected Lévy process. If V is a service time and λ is the arrival rate, then the Laplace exponent of the compound-Poisson net-input process is

$$\psi(s) = s - \lambda(1 - E[\exp(-sV)]) \ .$$

Example 5.2.2. *The gamma process.* A possible subordinator is the gamma process, which can be expressed via the Laplace exponent

$$\psi(s) = \int_0^\infty (e^{-sx} - 1)\frac{e^{-x/\eta}}{x}dx = -\log(1+\eta s)$$

for constant η; e.g., see p. 111 of Prabhu (1998). (The centering function is not needed in this case.) If we add a constant negative drift to the gamma process then we obtain a Lévy process with negative drift but without negative jumps, having Laplace exponent $\psi(s) = bs - \log(1+\eta s)$. If $b > \eta$, then $EL(1) < 0$ and we can apply Theorem 5.2.1. In this case, the steady-state ccdf H^c is easy to compute from its Laplace transform $H^c(s) = [1 - h(s)]/s$ by numerical inversion. The gamma process is a Lévy process without Brownian component; i.e., $b = \sigma^2 = 0$. The Lévy measure has density $\mu(dx) = x^{-1}e^{-x/\eta}$, $x > 0$. We can approximate the gamma process by a compound Poisson process by restricting μ to $[\epsilon, \infty)$ for some $\epsilon > 0$. ∎

For other properties of Lévy processes without negative jumps, see Takács (1967), Samorodnitsky and Taqqu (1994), Bertoin (1996) and Prabhu (1998). For a numerical inversion algorithm to calculate first-passage probabilities, see Rogers (2000).

5.3. A Fluid Queue Fed by On-Off Sources

This section is devoted to proving Theorem 8.3.1 in the book, which establishes a FCLT for the cumulative busy time of a single on-off source.

We first restate the theorem. Recall that $B_{n,i}$ is the i^{th} busy period and $I_{n,i}$ is the i^{th} idle period in the n^{th} model, in the sequence of models under consideration. Let

$$\begin{aligned}
\mathbf{B}_n(t) &\equiv c_n^{-1}\sum_{i=1}^{\lfloor nt \rfloor}(B_{n,i} - m_{B,n}) \\
\mathbf{I}_n(t) &\equiv c_n^{-1}\sum_{i=1}^{\lfloor nt \rfloor}(I_{n,i} - m_{I,n}) \\
\mathbf{N}_n(t) &\equiv c_n^{-1}[N_n(nt) - \gamma_n nt] \\
\mathbf{B}'_n(t) &\equiv c_n^{-1}[B_n(nt) - \xi_n nt], \quad t \geq 0\ , \qquad (3.1)
\end{aligned}$$

where again $\lfloor nt \rfloor$ is the integer part of nt,

$$\xi_n \equiv \frac{m_{B,n}}{m_{B,n} + m_{I,n}} \quad \text{and} \quad \gamma_n \equiv \frac{1}{m_{B,n} + m_{I,n}} \ . \tag{3.2}$$

We think of $m_{B,n}$ in (3.1) as the mean busy period, $EB_{n,i}$, and $m_{I,n}$ as the mean idle period, $EI_{n,i}$, in the case $\{(B_{n,i}, I_{n,i}) : i \geq 1\}$ is a stationary sequence for each n, but in general that is not required.

Theorem 5.3.1. (FCLT for the cumulative busy time) *If*

$$(\mathbf{B}_n, \mathbf{I}_n) \Rightarrow (\mathbf{B}, \mathbf{I}) \quad \textit{in} \quad (D, M_1)^2 \tag{3.3}$$

for $\mathbf{B}_n$ *and* $\mathbf{I}_n$ *in* (3.1), $c_n \to \infty$, $c_n/n \to 0$, $m_{B,n} \to m_B$, $m_{I,n} \to m_I$, *with* $0 < m_B + m_I < \infty$, *so that* $\xi_n \to \xi$ *with* $0 \leq \xi \leq 1$ *and* $\gamma_n \to \gamma > 0$ *for* ξ_n *and* γ_n *in* (3.2), *and*

$$P(Disc(\mathbf{B}) \cap Disc(\mathbf{I}) = \phi) = 1 \ , \tag{3.4}$$

then

$$(\mathbf{B}_n, \mathbf{I}_n, \mathbf{N}_n, \mathbf{B}'_n) \Rightarrow (\mathbf{B}, \mathbf{I}, \mathbf{N}, \mathbf{B}') \quad \textit{in} \quad (D, M_1)^4 \ , \tag{3.5}$$

for $\mathbf{N}_n$, $\mathbf{B}'_n$ *in* (3.1) *and*

$$\begin{aligned} \mathbf{N}(t) &\equiv -\gamma[\mathbf{B}(\gamma t) + \mathbf{I}(\gamma t)] \\ \mathbf{B}'(t) &\equiv (1 - \xi)\mathbf{B}(\gamma t) - \xi \mathbf{I}(\gamma t) \ . \end{aligned} \tag{3.6}$$

The possibility of the limit processes having discontinuous sample paths makes the required argument more complicated than what it might otherwise be. To make that clear, before presenting an argument that works, we present two false starts.

5.3.1. Two False Starts

For the first false start, note that the cumulative busy-time process can be bounded above and below by random sums by

$$c_n^{-1} \sum_{i=1}^{N_n(nt)} B_{n,i} \leq c_n^{-1} B_n(nt) \leq c_n^{-1} \sum_{i=1}^{N_n(nt)+1} B_{n,i} \ , \tag{3.7}$$

so let us start by trying to find limits for the outer terms in (3.7). We apply the continuous mapping theorem with addition (Section 12.7 in the

book) and the inverse map (Sections 13.7 and 13.8 in the book) to get, first, $\mathbf{B}_n + \mathbf{I}_n \Rightarrow \mathbf{B} + \mathbf{I}$ and then $\mathbf{N}_n \Rightarrow \mathbf{N}$ jointly.

As a consequence, we get $\mathbf{T}_n \Rightarrow \gamma e$, where

$$T_n(t) \equiv n^{-1} N_n(nt), \quad t \geq 0 \ . \tag{3.8}$$

Then we try to treat the term on the left in (3.7) by writing

$$\begin{aligned} & c_n^{-1} \left(\sum_{i=1}^{N_n(nt)} B_{n,i} - m_{n,2} \gamma_n nt \right) \\ = \ & c_n^{-1} \left(\sum_{i=1}^{\lfloor nt \rfloor} B_{n,i} - m_{n,2} \right) \circ \frac{N_n(nt)}{n} + m_{n,2} (c_n^{-1} [N_n(nt) - \gamma_n nt]) \\ \Rightarrow \ & \mathbf{B}(\gamma t) - m_2(\gamma[\mathbf{B}(\gamma t) + \mathbf{I}(\gamma t)]) = (1-\xi)\mathbf{B}(\gamma t) - \xi \mathbf{I}(\gamma t) \, . \end{aligned} \tag{3.9}$$

This argument works fine if $P(\mathbf{B} \in C) = 1$, but not otherwise. This argument is not valid here because we need to apply addition when the limit processes $\mathbf{B} \circ \gamma e$ and $-\gamma(\mathbf{B} \circ \gamma e + \mathbf{I} \circ \gamma e)$ typically have common discontinuities of opposite sign. (If they had the same sign, then we could apply Theorem 12.7.3 in the book.) Hence we need to find a different approach.

For our second false start, instead of (3.7), we find different bounds for the cumulative busy-time process, in particular, note that

$$\begin{aligned} \mathbf{B}'_n(t) &\leq c_n^{-1} \left[(1-\xi_n) \sum_{i=1}^{N_n(nt)+1} B_{n,i} - \xi_n \sum_{i=1}^{N_n(nt)} I_{n,i} \right] \\ &\leq c_n^{-1} \left[(1-\xi_n) \sum_{i=1}^{N_n(nt)+1} (B_{n,i} - m_{n,1}) - \xi_n \sum_{i=1}^{N_n(nt)} (I_{n,i} - m_{n,2}) \right] \\ &\quad + c_n^{-1} m_{n,1} \end{aligned}$$

$$\begin{aligned} \mathbf{B}'_n(t) &\geq c_n^{-1} \left[(1-\xi_n) \sum_{i=1}^{N_n(nt)} B_{n,i} - \xi_n \sum_{i=1}^{N_n(nt)+1} I_{n,i} \right] \\ &\geq c_n^{-1} \left[(1-\xi_n) \sum_{i=1}^{N_n(nt)} (B_{n,i} - m_{n,1}) - \xi_n \sum_{i=1}^{N_n(nt)+1} (I_{n,i} - m_{n,2}) \right] \\ &\quad - c_n^{-1} m_{n,2} \, . \end{aligned}$$

Note that the deterministic terms $c_n^{-1}m_{n,1}$ and $c_n^{-1}m_{n,2}$ are asymptotically negligible. Thus, let the asymptotically bounding processes be

$$\mathbf{B}_n^u(t) \equiv c_n^{-1}\left[(1-\xi_n)\sum_{i=1}^{N_n(nt)+1}(B_{n,i}-m_{n,1}) - \xi_n\sum_{i=1}^{N_n(nt)}(I_{n,i}-m_{n,2})\right] \tag{3.10}$$

and

$$\mathbf{B}_n^l(t) \equiv c_n^{-1}\left[(1-\xi_n)\sum_{i=1}^{N_n(nt)}(B_{n,i}-m_{n,1}) - \xi_n\sum_{i=1}^{N_n(nt)+1}(I_{n,i}-m_{n,2})\right]. \tag{3.11}$$

Also let

$$\mathbf{T}_n(t) \equiv \frac{N_n(nt)}{n}, \quad \mathbf{T}'_n(t) \equiv \frac{N_n(nt)+1}{t} \tag{3.12}$$

and

$$\mathbf{N}'_n(t) \equiv c_n^{-1}[N_n(nt)+1-\gamma_n nt], \quad t \geq 0\,. \tag{3.13}$$

As before, we apply the continuous mapping theorem with the addition and the inverse map to get, first $\mathbf{B}_n + \mathbf{I}_n \Rightarrow \mathbf{B} + \mathbf{I}$ and then $\mathbf{N}_n \Rightarrow \mathbf{N}$ and $\mathbf{N}'_n \Rightarrow \mathbf{N}$, all jointly. Given $\mathbf{N}_n \Rightarrow \mathbf{N}$ and $\mathbf{N}'_n \Rightarrow \mathbf{N}$ we obtain $\mathbf{T}_n \Rightarrow \gamma e$ and $\mathbf{T}'_n \Rightarrow \gamma e$ by multiplying by c_n/n. Applying the composition map, we obtain

$$\mathbf{B}_n^u = (1-\xi_n)\mathbf{B}_n \circ \mathbf{T}'_n - \xi_n \mathbf{I}_n \circ \mathbf{T}_n \Rightarrow \mathbf{B}' \tag{3.14}$$

and

$$\mathbf{B}_n^l = (1-\xi_n)\mathbf{B}_n \circ \mathbf{T}_n - \xi_n \mathbf{I}_n \circ \mathbf{T}'_n \Rightarrow \mathbf{B}'\,, \tag{3.15}$$

again jointly with the other limits. Hence we are close to obtaining (3.5). However, even though $(\mathbf{B}_n^l, \mathbf{B}_n^u) \Rightarrow (\mathbf{B}', \mathbf{B}')$ and $\mathbf{B}_n^l \leq \mathbf{B}'_n \leq \mathbf{B}_n^u$, we cannot deduce that $\mathbf{B}'_n \Rightarrow \mathbf{B}'$ in (D, M_1).

5.3.2. The Proof

We can deduce that $\mathbf{B}'_n \Rightarrow \mathbf{B}'$ in the weaker Skorohod M_2 topology by this reasoning, though, by virtue of Corollary 12.11.4 in the book, from which we can deduce convergence of the finite-dimensional distributions. To get the desired M_1 limit, it thus suffices to apply Theorem 12.5.1 (iv) in the book and control the oscillations as in equation (12.5.3) of the book. To do

so, we introduce a slightly different approximation. Let

$$\mathbf{B}_n^a(t) = c_n^{-1} \left[(1-\xi_n) \sum_{i=1}^{N_n^B(nt)} (B_{n,i} - m_{n1}) - \xi_n \sum_{i=1}^{N_n^I(nt)} (I_{n,i} - m_{n,2}) \right], \tag{3.16}$$

where $N_n^B(t)$ and $N_n^I(t) = N_n(t)$ are the number of complete busy periods and idle periods by time t. Reasoning as with (3.14) and (3.15) we can deduce that $\mathbf{B}_n^a \Rightarrow \mathbf{B}'$. However, we can make a stronger connection between $\mathbf{B}_n^a$ and $\mathbf{B}_n$. Note that

$$N_n^I(t) = N_n(t) \le N_n^B(t) \le N_n(t) + 1$$

and

$$\mathbf{B}_n^a \circ \mathbf{S}_n = \mathbf{B}_n \circ \mathbf{S}_n \quad \text{and} \quad \mathbf{B}_n^a \circ \mathbf{S}_n' = \mathbf{B}_n \circ \mathbf{S}_n'$$

where

$$\mathbf{S}_n(t) \equiv n^{-1}\tau_{n,\lfloor nt \rfloor}, \quad \mathbf{S}_n'(t) \equiv n^{-1}\tau'_{n,\lfloor nt \rfloor},$$

$\tau_{n,0} = 0$,

$$\tau_{n,k} \equiv B_{n,1} + I_{n,1} + \cdots + B_{n,k} + I_{n,k}, \quad k \ge 1,$$

and

$$\tau'_{n,k} \equiv \tau_{n,k} + B_{n,k+1}, \quad k \ge 0.$$

Moreover $\mathbf{B}_n^a$ is piecewise-constant and $\mathbf{B}_n$ is piecewise linear in each of the intervals $[n^{-1}\tau_{n,k}, n^{-1}\tau'_{n,k}]$ and $[n^{-1}\tau'_{n,k}, n^{-1}\tau_{n,k+1}]$. Hence we can relate the oscillation of $\mathbf{B}_n$ to those of $\mathbf{B}_n^a$.

First, we can apply the Skorohod representation theorem to replace convergence in distribution by convergence w.p.1. We obtain $\mathbf{B}_n^a \to \mathbf{B}'$ w.p.1 for new versions of these processes. From the specific structure above, we can construct the corresponding special version of $\mathbf{B}_n'$ associated with $\mathbf{B}_n^a$. (It is the piecewise-linear interpolation of the piecewise-constant function.) Since $\mathbf{B}_n^a \to \mathbf{B}'$, $\mathbf{S}_n \to \gamma^{-1}e$ and $\mathbf{S}_n' \to \gamma^{-1}e$ for the new versions, we can deduce that $\mathbf{B}_n'(t) \to \mathbf{B}'(t)$ w.p.1 for each continuity point t of $\mathbf{B}'$. (We also got this part from the convergence of $\mathbf{B}_n^l$ and $\mathbf{B}_n^u$.) Let w_s be the M_1 oscillation function over the interval $[0,T]$, where T is chosen to be a continuity point of $\mathbf{B}'$, i.e.,

$$w_s(x,\delta) \equiv \sup_{0 \vee (t-\delta) \le t_1 < t_2 < t_3 \le (t+\delta) \wedge T} \{|x(t_2) - [x(t_1), x(t_3)]|\}$$

where $[x(t_1), x(t_3)]$ is the line segment connecting $x(t_1)$ and $x(t_2)$. From the properties above, we can deduce that

$$w_s(\mathbf{B}_n', \delta) \le w_s(\mathbf{B}_n^a, 2\delta)$$

for all suitably large n. Since $\mathbf{B}_n^a \to \mathbf{B}'$, we deduce that

$$\lim_{\delta \downarrow 0} \overline{\lim_{n \to \infty}} \; w_s(\mathbf{B}_n^a, \delta) = 0\,, \tag{3.17}$$

which implies the same limit with $\mathbf{B}_n^a$ replaced by $\mathbf{B}_n'$ in (3.17). By the characterization of M_1 convergence in Theorem 12.5.1 (iv) in the book, we get $\mathbf{B}_n' \to \mathbf{B}'$ w.p.1 (in D, M_1) for the special versions and thus $\mathbf{B}_n' \Rightarrow \mathbf{B}'$ for the original versions. This can be done jointly with the other processes, so that we get (3.5).

5.4. From Queue Lengths to Waiting Times

In this section, following Puhalskii (1994), we show how the continuous-mapping approach with the inverse map and nonlinear centering term, Theorem 13.7.4 in the book, can be used to convert limits for arrival, departure and queue-length processes into associated limits for waiting-time and workload processes in quite general queueing models. The nonlinear centering enables us to capture nonstationary phenomena.

5.4.1. The Setting

The setting is a family of queueing models indexed by n. Suppose that all arrivals eventually get served and then depart, so that the queue length (number of customers in the system) at time t is just the initial queue length plus the arrivals minus the departures, i.e.,

$$Q_n(t) = Q_n(0) + A_n(t) - D_n(t), \quad t \geq 0\,, \tag{4.1}$$

where $Q_n(t)$ is the queue length at time t, $A_n(t)$ is the number of arrivals in the interval $[0, t]$, and $D_n(t)$ is the number of departures in the interval $[0, t]$, all in model n. To treat customer waiting times (but not the workload), we need to make assumptions about the service mechanism. In particular, we assume that the customers are served one at a time in order of their arrival. Thus, we are again in the setting of the standard single-server queue. Let $A_n(t)$ count the new arrivals, and let $D_n(t)$ counts all departures, including those customers originally in the system at time 0. Note that $\{A_n(t) : t \geq 0\}$ and $\{D_n(t) : t \geq 0\}$ are counting processes. As a regularity condition, we assume that $A_n(0) = D_n(0) = 0$.

5.4.2. The Inverse Map with Nonlinear Centering

We can use the inverse map to define related quantities of interest. Let $A_{n,k}$ be the arrival time of the k^{th} arriving customer, $D_{n,k}$ the departure time of the k^{th} arriving customer and $L_n(t)$ the workload facing the server at time t, not counting arrivals after time t (the virtual waiting time), all in model n. Then

$$\begin{aligned} A_{n,k} &\equiv \inf\{s \geq 0 : A_n(s) > (k-1)^+\} , \\ D_{n,k} &\equiv \inf\{s \geq 0 : D_n(s) > (Q_n(0) + k - 1)^+\} , \\ L_n(t) &\equiv \inf\{s \geq 0 : D_n(s) > Q_n(0) + A_n(t)\} \end{aligned} \tag{4.2}$$

for $k \geq 1$ and $t \geq 0$, where $(x)^+ = \max\{x, 0\}$.

Let $W_{n,k}$ be the waiting time for arriving customer k to begin service and let $W'_{n,k}$ be the waiting time until customer k completes service. Then, under the assumptions about the service mechanism above,

$$W_{n,k} \equiv [D_{n,k-1} - A_{n,k}]^+ , \tag{4.3}$$

and

$$W'_{n,k} \equiv D_{n,k} - A_{n,k}, \quad k \geq 1 . \tag{4.4}$$

Suppose that the time scaling is already incorporated in the models indexed by n. We assume that functional weak laws of large numbers (FWLLNs) holds with additional space scaling by n and that FCLTs hold with additional space scaling by c_n after centering. Thus, let

$$\begin{aligned} \hat{\mathbf{X}}_n(t) &\equiv n^{-1} D_n(t), \\ \hat{\mathbf{Y}}_n(t) &\equiv n^{-1} A_n(t), \\ \hat{\mathbf{Q}}_n(t) &\equiv n^{-1} Q_n(t), \\ \mathbf{X}_n(t) &\equiv c_n(\hat{\mathbf{X}}_n - \mathbf{x}), \\ \mathbf{Y}_n(t) &\equiv c_n(\hat{\mathbf{Y}}_n - \mathbf{y}), \\ \mathbf{Q}_n(t) &\equiv c_n(\hat{\mathbf{Q}}_n - \mathbf{q}), \quad t \geq 0 . \end{aligned} \tag{4.5}$$

We assume that

$$(\hat{\mathbf{X}}_n, \hat{\mathbf{Y}}_n, \hat{\mathbf{Q}}_n) \Rightarrow (\mathbf{x}, \mathbf{y}, \mathbf{q}) \quad \text{in} \quad (D^3, WM_1) \tag{4.6}$$

where $\mathbf{x}, \mathbf{y} \in D_\uparrow$, $\mathbf{q} \in D$ and, by (4.1),

$$\mathbf{q}(t) = \mathbf{q}(0) + \mathbf{y}(t) - \mathbf{x}(t), \quad t \geq 0 . \tag{4.7}$$

We will also impose smoothness conditions on $\mathbf{x}$ and $\mathbf{y}$. In addition, we assume that $c_n \to \infty$ and

$$(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Q}_n) \Rightarrow (\mathbf{X}, \mathbf{Y}, \mathbf{Q}) \quad \text{in} \quad (D^3, WM_1)\,. \tag{4.8}$$

As a consequence of (4.1) and (4.5)–(4.8),

$$\mathbf{Q}(t) - \mathbf{Q}(0) = \mathbf{A}(t) - \mathbf{D}(t) \quad \text{for} \quad t > 0\,. \tag{4.9}$$

Given the FWLLN (4.6) and the FCLT (4.8), we want to establish related limits for appropriately scaled versions of the random variables $A_{n,k}$, $D_{n,k}$, $L_n(t)$, $W_{n,k}$ and $W'_{n,k}$ in (4.2)–(4.4). For that purpose, let

$$\hat{\mathbf{D}}_n(t) \equiv D_{n,\lfloor nt \rfloor}, \quad \hat{\mathbf{A}}_n(t) \equiv A_{n,\lfloor nt \rfloor}, \quad \hat{\mathbf{L}}_n(t) = L_n(t) \tag{4.10}$$

and

$$\hat{\mathbf{W}}_n(t) \equiv W_{n,\lfloor nt \rfloor} \quad \text{and} \quad \hat{\mathbf{W}}'_n(t) \equiv W'_{n,\lfloor nt \rfloor},\ t \geq 0\,. \tag{4.11}$$

We now form the final scaled random elements of D. Let

$$\begin{aligned}
\mathbf{U}_n(t) &\equiv c_n(\hat{\mathbf{X}}_n^{-1} - \mathbf{x}^{-1}),\\
\mathbf{V}_n(t) &\equiv c_n(\hat{\mathbf{Y}}_n^{-1} - \mathbf{y}^{-1}),\\
\mathbf{A}_n(t) &\equiv c_n(\hat{\mathbf{A}}_n - \mathbf{y}^{-1}),\\
\mathbf{D}_n(t) &\equiv c_n(\hat{\mathbf{D}}_n - \mathbf{x}^{-1} \circ \mathbf{z}_1),\\
\mathbf{L}_n(t) &\equiv c_n(\hat{\mathbf{L}}_n - \mathbf{x}^{-1} \circ \mathbf{z}_2),\\
\mathbf{W}_n(t) &\equiv c_n(\hat{\mathbf{W}}_n - (\mathbf{x}^{-1} \circ \mathbf{z}_1 - \mathbf{y}^{-1}),\\
\mathbf{W}'_n(t) &\equiv c_n(\hat{\mathbf{W}}'_n - (\mathbf{x}^{-1} \circ \mathbf{z}_1 - \mathbf{y}^{-1}), \quad t \geq 0\,.
\end{aligned} \tag{4.12}$$

We now state the theorem.

Theorem 5.4.1. (FCLT for the workload and waiting time given a FCLT for arrivals, departures and queue length) *Suppose that the limit* (4.8) *holds for* $\mathbf{X}_n$, $\mathbf{Y}_n$, $\mathbf{Q}_n$ *in* (4.5), *where* $c_n \to \infty$, $\mathbf{x}, \mathbf{y} \in \Lambda$ *and are absolutely continuous with continuous positive derivatives* $\dot{\mathbf{x}}$, $\dot{\mathbf{y}}$, *and* $P(\mathbf{X}(0) = 0) = P(\mathbf{Y}(0) = 0) = 1$. *Then, jointly with* (4.8),

$$(\mathbf{U}_n, \mathbf{V}_n, \mathbf{A}_n, \mathbf{D}_n) \Rightarrow (\mathbf{U}, \mathbf{V}, \mathbf{A}, \mathbf{D}) \tag{4.13}$$

in (D^4, WM_1) *for* $\mathbf{U}_n$, $\mathbf{V}_n$, $\mathbf{A}_n$ *and* $\mathbf{D}_n$ *in* (4.12), *where*

$$\mathbf{U} = \frac{-\mathbf{X} \circ \mathbf{x}^{-1}}{\dot{\mathbf{x}} \circ \mathbf{x}^{-1}}, \quad \mathbf{V} = \mathbf{A} = \frac{-\mathbf{Y} \circ \mathbf{y}^{-1}}{\dot{\mathbf{y}} \circ \mathbf{y}^{-1}} \tag{4.14}$$

and

$$\mathbf{D} = \frac{-\mathbf{X} \circ \mathbf{x}^{-1} \circ \mathbf{z}_1 + \mathbf{Q}(0)\mathbf{1}}{\dot{\mathbf{x}} \circ \mathbf{x}^{-1} \circ \mathbf{z}_1}, \quad \mathbf{z}_1 = \mathbf{q}(0)\mathbf{1} + \mathbf{e}\,, \tag{4.15}$$

where $\mathbf{e}(t) = t$ *for* $t \geq 0$. *If, in addition,*

$$P(Disc(\mathbf{X} \circ \mathbf{x}^{-1} \circ \mathbf{z}_2) \cap Disc(\mathbf{Y}) = \phi) = 1 \tag{4.16}$$

for

$$\mathbf{z}_2 = \mathbf{q}(0)\mathbf{1} + \mathbf{y}\ , \tag{4.17}$$

then, jointly with (4.8) and (4.13),

$$\mathbf{L}_n \Rightarrow \mathbf{L} \tag{4.18}$$

for $\mathbf{L}_n$ *in* (4.12), *where*

$$\mathbf{L} = \frac{-\mathbf{X} \circ \mathbf{x}^{-1} \circ \mathbf{z}_2 + \mathbf{Y} + \mathbf{Q}(0)\mathbf{1}}{\dot{\mathbf{x}} \circ \mathbf{x}^{-1} \circ \mathbf{z}_2}\ . \tag{4.19}$$

If, in addition,

$$P(Disc(\mathbf{A}) \cap Disc(\mathbf{D}) = \phi) = 1\ , \tag{4.20}$$

then, jointly with (4.8), (4.13) *and* (4.18),

$$(\mathbf{W}_n, \mathbf{W}_n') \Rightarrow (\mathbf{D} - \mathbf{A}, \mathbf{D} - \mathbf{A}) \tag{4.21}$$

in (D^2, WM_1) *for* $\mathbf{W}_n$ *and* $\mathbf{W}_n'$ *in* (4.12).

In preparation for the proof, we now restate Theorem 13.7.4 from the book. Recall that $D_\uparrow$ is the subset of all nondecreasing nonnegative functions in D. Recall that D_u is the subset of all functions in $D([0,\infty), \mathbb{R})$ that are unbounded above and satisfy $x(0) \geq 0$.

The following is Puhalskii's (1994) result extended to allow discontinuous limits.

Theorem 5.4.2. *Suppose that* $x_n \in D_u$, $y_n \in D_\uparrow$, $c_n \to \infty$,

$$c_n(x_n - x, y_n - y) \to (u, v) \quad in \quad D \times D \tag{4.22}$$

with one of the J_1, M_1 *or* M_2 *topologies, where* $u(0) = 0$, u *has no positive jumps if the topology is* J_1,

$$Disc(u \circ x^{-1} \circ y) \cap Disc(v) = \phi\ , \tag{4.23}$$

$y \in C_{\uparrow\uparrow}$ *and* x *is absolutely continuous with a continuous positive derivative* $\dot{x}$, *then*

$$c_n(x_n^{-1} \circ y_n - x^{-1} \circ y) \to \frac{v - u \circ x^{-1} \circ y}{\dot{x} \circ x^{-1} \circ y} \quad in \quad D \tag{4.24}$$

with the same topology.

Proof of Theorem 5.4.1. We start by applying the Skorohod representation theorem to replace convergence in distribution by convergence w.p.1. For simplicity, we do not introduce new notation for these special versions of the random functions converging w.p.1. Thus consider a single sample path for which the limit (4.8) holds. Now we can apply the deterministic convergence-preservation results. From (4.2)–(4.11), we see that we can represent $\hat{\mathbf{D}}_n$, $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{L}}_n$ in terms of $\hat{\mathbf{X}}_n$ and $\hat{\mathbf{Y}}_n$ via the inverse map

$$\begin{aligned} \hat{\mathbf{A}}_n(t) &\equiv \inf\{s \geq 0 : A_n(s) > \lfloor nt \rfloor - 1)\} \\ &= \inf\{s \geq 0 : \hat{\mathbf{Y}}_n(s) > (\lfloor nt \rfloor - 1)/n\} \\ &= (\hat{\mathbf{Y}}_n^{-1} \circ \xi_n)(t), \quad t \geq 0\,, \end{aligned} \tag{4.25}$$

where

$$\xi_n(t) = (\lfloor nt \rfloor - 1)^+/n, \quad t \geq 0\,, \tag{4.26}$$

$$\begin{aligned} \hat{\mathbf{D}}_n(t) &\equiv \inf\{s \geq 0 : D_n(s) > (Q_n(0) + \lfloor nt \rfloor - 1)^+\} \\ &= \inf\{s \geq 0 : \hat{\mathbf{X}}_n(s) > \{Q_n(0) + \lfloor nt \rfloor - 1)^+/n\} \\ &= (\hat{\mathbf{X}}_n^{-1} \circ \zeta_n)(t), \quad t \geq 0\,, \end{aligned} \tag{4.27}$$

where

$$\zeta_n(t) = (Q_n(0) + \lfloor nt \rfloor - 1)^+/n, \quad t \geq 0\,, \tag{4.28}$$

and

$$\begin{aligned} \hat{\mathbf{L}}_n(t) &\equiv \inf\{s \geq 0 : D_n(s) > Q_n(0) + A_n(nt)\} \\ &= \inf\{s \geq 0 : \hat{\mathbf{X}}_n(s) > \hat{\mathbf{Q}}_n(0) + \hat{\mathbf{Y}}_n(t)\} \\ &= [\hat{\mathbf{X}}_n^{-1} \circ (\hat{\mathbf{Q}}_n(0)\mathbf{1} + \hat{\mathbf{Y}}_n)](t), \quad t \geq 0 \end{aligned} \tag{4.29}$$

where $\mathbf{1}(t) \equiv 1$, $t \geq 0$. Given (4.3)–(4.27),

$$\hat{\mathbf{W}}_n(t) = [(\hat{\mathbf{D}}_n \circ \xi_n)(t) - \hat{\mathbf{A}}_n(t)]^+, \quad t \geq 0\ , \tag{4.30}$$

for ξ_n in (4.26) and

$$\hat{\mathbf{W}}_n'(t) = (\hat{\mathbf{D}}_n - \hat{\mathbf{A}}_n)(t), \quad t \geq 0\ . \tag{4.31}$$

We now return to the proof of (4.13). First, for the inverse processes $\hat{\mathbf{X}}_n^{-1}$ and $\hat{\mathbf{Y}}_n^{-1}$, we apply Theorem 13.7.2 from the book. Given those two limits, we treat $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{D}}_n$ by applying the composition result, Theorem 12.3.1. Alternatively, we directly apply Theorem 5.4.2 above, noting that $\xi_n \to \mathbf{e}$,

$\zeta_n \to \mathbf{z}_1$, $c_n(\xi_n - e) \to \mathbf{0}$ and $c_n(\zeta_n - z_1) \Rightarrow \hat{\mathbf{Q}}(0)\mathbf{1}$. To treat $\hat{\mathbf{L}}_n$ we again apply Theorem 12.3.1 or Theorem 5.4.2 above, using the fact that $\hat{\mathbf{Q}}_n(0)\mathbf{1} + \hat{\mathbf{Y}}_n \to \mathbf{z}_2$ and $c_n(\hat{\mathbf{Q}}_n(0)\mathbf{1} + \hat{\mathbf{Y}}_n - \mathbf{z}_2) \to \mathbf{Y} + \mathbf{Q}(0)\mathbf{1}$ in D. Finally, to treat $\hat{\mathbf{W}}_n$ and $\hat{\mathbf{W}}_n^{'}$, we use the subtraction map. We first apply subtraction directly to $\hat{\mathbf{W}}_n^{'}$ in (4.31). Since $\xi_n \Rightarrow \mathbf{e}$, we can conclude that $\mathbf{W}_n$ has the same limit as $\mathbf{W}'_n$. ∎

Remark 5.4.1. If Theorem 5.4.1 holds for stationary models, then $\mathbf{x} = \mathbf{y} = \lambda\mathbf{e}$, and $\mathbf{q} = \mathbf{q}(0)\underline{1}$. Suppose in addition that $\mathbf{q}(0) = 0$. By (4.9), if we cannot conclude that the limit processes almost surely have continuous paths, then we should anticipate $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Q}$ can have common discontinuities. Then

$$\mathbf{U} = -\lambda^{-1}\mathbf{X} \circ \lambda^{-1}\mathbf{e} \tag{4.32}$$

and

$$\mathbf{V} = \mathbf{A} = -\lambda^{-1}\mathbf{Y} \circ \lambda^{-1}\mathbf{e}\,. \tag{4.33}$$

Condition (4.16) then becomes

$$P(Disc(\mathbf{X}) \cap Disc(\mathbf{Y}) = \phi) = 1 \tag{4.34}$$

and

$$\mathbf{D} = \lambda^{-1}(\mathbf{Y} - \mathbf{X} + \mathbf{Q}(0)\mathbf{1}) = \lambda^{-1}\mathbf{Q}\,. \tag{4.35}$$

Then the centering terms in (4.21) become

$$\mathbf{x}^{-1} \circ \mathbf{z}_1 - \mathbf{y}^{-1} = \lambda^{-1}\mathbf{e} - \lambda^{-1}\mathbf{e} = \mathbf{0} \tag{4.36}$$

and

$$\begin{aligned} \mathbf{D} - \mathbf{A} &= \lambda^{-1}(\mathbf{Y} - \mathbf{X} + \mathbf{Q}(0)\mathbf{1}) + \lambda^{-1} \circ \mathbf{X} \circ \lambda^{-1}\mathbf{e} \\ &= \lambda^{-1}(\mathbf{Q} + \mathbf{X} \circ \lambda^{-1}\mathbf{e})\,. \end{aligned}$$

∎

5.4.3. An Application to Central-Server Models

Following Puhalskii (1994), we illustrate how Theorem 5.4.1 can be applied by considering a limit for a central-server model. Central-server models were originally introduced to model the contention among programs for the processor and input-output devices in a multiprogrammed computer system; e.g., see Section 3.4.2 of Lavenberg and Sauer (1983). The specific model we consider is a closed queueing network with $n + 1$ single-server queues,

one of which is called the central-server queue while the others are called peripheral queues. There are n customers (jobs) in the network, one for each peripheral queue. Each customer has a designated distinct peripheral queue. Each customer circulates between the central-server queue and its own designated peripheral queue. The customers are served one at a time in order of arrival at the central-server queue. The service times are assumed to be mutually independent exponential random variables. (That ensures that the closed network has a product-form steady-state distribution.) Let the mean service time at each peripheral queue be λ^{-1}, and let the mean service time at the central-server queue be $(n\mu)^{-1}$.

Since only one customer receives service at each peripheral queue, there is no contention there. Thus, each customer enters service at its peripheral immediately upon arrival. Consequently, the $(n+1)$-queue model is equivalent to a 2-queue model, with one queue being the central-server queue and the other queue being an infinite-server queue. Moreover, the number of customers at the central-server queue evolves as a birth-and-death process with state-dependent transition rates. Let $Q_n(t)$ denote the number of customers at the central-server queue at time t, as a function of n. When $Q_n(t) = k$, the birth (arrival) rate is $(n-k)\lambda$ and the death (service) rate is $n\mu$. Hence the steady-state distribution is easy to calculate.

However, it is also of interest to consider limits as $n \to \infty$ in order to better understand the behavior of such systems with fast central servers and many customers. First a FLLN is quite elementary. For that purpose, let $A_n(t)$ and $D_n(t)$ count the numbers of arrivals and departures, respectively, at the central-server queue in the interval $[0,t]$. Then form the scaled processes $\hat{\mathbf{X}}_n$, $\hat{\mathbf{Y}}_n$ and $\hat{\mathbf{Q}}_n$ as in (4.5). It is then relatively elementary to show that, if $\hat{\mathbf{Q}}_n(0) = \mathbf{q}(0)$, $0 \le q(0) \le 1$, then the FWLLN in (4.6) holds here with

$$\mathbf{x}(t) = \mu t, \quad \mathbf{y}(t) = \lambda \int_0^t [1 - \mathbf{q}(s)]ds \tag{4.37}$$

and $\mathbf{q}$ satisfying the ordinary differential equation

$$\dot{\mathbf{q}}(t) \equiv \frac{d\mathbf{q}}{dt}(t) = \lambda(1 - \mathbf{q}(t)) - \mu \ . \tag{4.38}$$

Kogan, Lipster and Smorodinskii (1986) then established the following result; also see Chapter 8, Section 3, of Liptser and Shiryaev (1989) and Puhalskii (1994).

Theorem 5.4.3. (FCLT for the central-server model) *If*

$$\sqrt{n}[\mathbf{Q}_n(0) - q(0)] \Rightarrow \mathbf{Q}(0) \quad in \quad \mathbb{R}, \tag{4.39}$$

then the joint limit (4.8) *holds with*

$$\mathbf{X}(t) = \sqrt{\mu}\mathbf{B}_2(t) \ , \tag{4.40}$$

$$\mathbf{Y}(t) = \int_0^t \sqrt{\lambda(1-\mathbf{q}(s)}d\mathbf{B}_1(s) - \lambda\int_0^t \mathbf{Q}(s)ds \tag{4.41}$$

and

$$\mathbf{Q}(t) = \mathbf{Q}(0) + \mathbf{X}(t) - \mathbf{Y}(t), \quad t \geq 0 \ . \tag{4.42}$$

The limit process $\mathbf{Q}$ *can be expressed as the solution to*

$$\begin{aligned} \mathbf{Q}(t) = \mathbf{Q}(0) \quad &- \quad \lambda\int_0^t \mathbf{Q}(s)ds \\ &+ \quad \int_0^t \sqrt{\lambda(1-\mathbf{q}(s))}d\mathbf{B}_1(s) - \sqrt{\mu}\mathbf{B}_2(t) \, . \end{aligned} \tag{4.43}$$

We can now combine Theorems 5.4.1 and 5.4.3 to obtain associated limits for the scaled versions of $\hat{\mathbf{A}}_n$, $\hat{\mathbf{D}}_n$, $\hat{\mathbf{L}}_n$ in (4.10) and $\hat{\mathbf{W}}_n$ and $\hat{\mathbf{W}}_n'$ in (4.11), as stated in Theorem 5.4.1. Theorem 5.4.1 is genuinely helpful here, because these limits are not so easy to obtain directly.

Theorem 5.4.1 has also been applied by Mandelbaum, Massey, Reiman and Stolyar (1999).