3

Hierarchical Modeling of Stochastic Networks, Part II: Strong Approximations

Hong Chen
Avi Mandelbaum

3.1 Introduction

The goal in this part is to establish strong approximations for a family of open queueing networks. We cover, in particular, nonparametric Jackson networks. These are the classical open Jackson queueing networks, but without the parametric assumptions of exponential interarrival and service times (Section 2.5.2 of Part I). The basic results are Functional Strong Approximations (FSAT, Theorem 3.4.1) and a Functional Law of Iterated Logarithm (FLIL, Theorem 3.4.2). These readily imply fluid approximations, formalized by a Functional Strong Law of Large Numbers (FSLLN, Corollary 3.4.3), and diffusion approximations, formalized by a Functional Central Limit Theorem (FCLT, Corollary 3.4.4).

Our approach entails representing performance measures of interest, for example queue-length, as transformations of primitives, for example interarrival and service times. These transformations turn out to be Lipschitz continuous. Thus limit theorems and approximations of the primitives, specifically FSAT's, FSLLN's and FCLT's, carry over to the desired performance measures, and with the same order of error.

In the literature, the prevalent justifications for fluid and diffusion approximations have been FSLLN's and FCLT's. Strong approximations provide a framework which is conceptually different and, whenever applicable, it is in our opinion also superior: the framework is simple and direct - one first derives strong approximations, without any explicit rescaling; it unifies the classical FSLLN's, FLIL's and FCLT's - and these theorems, which are magnifying by-products of the approximations, reveal further insight; the framework also enables quantification of errors, in a way that is visibly monotone in the assumptions on the primitives - as moments of higher order are assumed finite, the better the approximation; finally, the mathematical toll for making strong approximations rigorous is more-than-minimal

moment-conditions on the primitives - but as far as current applications are concerned, this toll seems negligible.

3.2 The Model

3.2.1 Primitives and Dynamics

Our queueing network consists of K service stations, indexed by $k = 1, \cdots, K$. Each station k constitutes a server, called server k, and a queue, called queue k. Server k is dedicated to serving customers waiting in queue k. After being served, customers either leave the network or rejoin one of its queues in anticipation of additional service.

The network's dynamics are described in terms of the following vector primitives, the coordinates of which are all integer-valued: a K-dimensional nonnegative vector $Z(0)$, and a sequence of K-dimensional RCLL vector processes $F^k = \{F^k(t), t \geq 0\}$, $k = 0, 1, ..., K$. The kth coordinate of $Z(0)$, $Z_k(0)$, represents the number of jobs initially at station k. The kth coordinate of $F^0(t)$, $F^0_k(t)$, indicates the accumulated number of exogenous arrivals to station k up to time t. For $j, k = 1, ..., K$, with $j \neq k$, the jth coordinate of $F^k(s)$, $F^k_j(s)$, models the number of service completions at station k, which switch directly to station j during its first s units of busy-time; the negative of the kth coordinate, $-F^k_k(s)$, $k = 1, ..., K$, stands for the total number of departures (service completions minus immediate feedbacks) from station k during its first s units of busy time.

The sample paths of the *flow* processes F^0_j, F^k_j, $j \neq k$, and $-F^k_k$ are assumed to be nondecreasing, all with $F^k(0) = 0$, $k = 0, \ldots, K$. Adding the assumption that $-F^k_k$ has only unit jumps suffices to guarantee existence of the *queue length* process $Z = \{Z(t), t \geq 0\}$ and the *busy time* process $B = \{B(t), t \geq 0\}$. They are implicitly defined as the unique solutions to the *flow-balance* relation

$$Z(t) = Z(0) + F^0(t) + \sum_{k=1}^{K} F^k[B_k(t)], \tag{3.1}$$

subject to the *work-conserving* constraints

$$B_k(t) = \int_0^t 1[Z_k(s) > 0]ds, \qquad k = 1, ..., K \tag{3.2}$$

(Note that $Z(t) \geq 0$ must hold at all $t \geq 0$.)

3.2.2 Underlying Assumptions and Parameters

The primitives $Z(0)$ and F^k, $k = 0, \ldots, K$, are defined on a common probability space and, strictly for convenience, they are taken to be mutually independent. For $k = 0, \ldots, K$, we assume that there exist K-dimensional nonnegative vectors α^k, $K \times K$ covariance matrices Γ^k, K-dimensional mutually independent standard Wiener processes $W^k = \{W^k(t), t \geq 0\}$ such that

$$FSAT: \quad \sup_{0 \leq t \leq T} |F^k(t) - \alpha^k t - (\Gamma^k)^{1/2} W^k(t)| = o(T^{1/r}), \quad \text{as } T \uparrow \infty, \tag{3.3}$$

for some scalar $r > 2$. It is further assumed that the $K \times K$ matrix $[\alpha^1, ..., \alpha^K]$ has the form

$$[\alpha^1, ..., \alpha^K] = [P' - I]\text{diag}(\mu), \tag{3.4}$$

where P is a $K \times K$ substochastic matrix with spectral radius less than unity, and μ is a positive K-dimensional vector.

One interprets α^0_k, the kth coordinate of α^0, as the long-run average rate of exogenous arrivals to station k; μ_k, the kth coordinate of μ, as the long-run average potential rate of service completions from station k; out of these completions, a long-run average fraction p_{jk}, the (j,k)th component of P, switch directly to station k. We refer to α^0 and μ as the *arrival* and *service* rates, and to P as the *transition* matrix.

For later use we record that Strassen's FLIL for the Brownian motion (prior to (1.7) in Part I), applied to W^k in (3.3), yields

$$FLIL: \quad \sup_{0 \leq t \leq T} |F^k(t) - \alpha^k t| = O(\sqrt{T \text{log} \text{log} T}), \quad \text{as } T \uparrow \infty, \tag{3.5}$$

for $k = 0, ..., K$.

3.2.3 Nonparametric Jackson Networks

Our framework, in particular FSAT (3.3) and FLIL (3.5), covers nonparametric open Jackson queueing networks (Section 2.5.2 in Part I). Here the flows are constructed from lower-level primitives, which constitute the following mutually independent K-dimensional entities: an arrival process A, a service process S and a routing sequence $\xi^k = \{\xi^k(\ell), \ell = 1, 2, \cdots\}$. The kth component of A, $A_k = \{A_k(t), t \geq 0\}$, is a renewal process that models exogenous arrivals to station k. The kth component of S, $S_k = \{S_k(u), u \geq 0\}$, is a renewal process that models service completions from station k: $S_k(u)$ represents the number of service completions by server k during its first u units of busy-time. Finally, $\xi^k = \{\xi^k(\ell), \ell = 1, 2, \cdots\}$ models the routing mechanism enforced at station k: its jth component $\xi^k_j(\ell)$ is the indicator of the event that the ℓth customer served at station k, upon completion of its service, is routed directly to queue j, $j = 1, \cdots, K$. (Formally, $\xi^k_j(\ell)$ equals 1 when the event that it indicates occurs and 0 when it does not.) To ease the notation introduce

$$R^k(n) = \sum_{\ell=1}^{n} [\xi^k(\ell) - e^k], \quad n = 0, 1, 2, \ldots, \tag{3.6}$$

for $k = 1, \ldots, K$, and denote its coordinates by $R_j^k(t)$, $j, k = 1, ..., K$, e^k is the K-dimensional kth unit vector, $k = 1, \ldots, K$. The connection with the model in Section 3.2.1 is revealed through

$$F^0(t) = A(t), \tag{3.7}$$
$$F^k(t) = R^k[S_k(t)], \quad k = 1, ..., K, \tag{3.8}$$

at all $t \geq 0$.

Consider station k, $k = 1, \ldots, K$. We assume that the renewal processes A_k and S_k are constructed from i.i.d. intervals with finite moments of order r, for some $r > 2$. Then FSAT's can be proved, that establish a probability space, supporting independent standard Wiener processes W^k, $k = 0, ..., K$, for which (3.3) holds. The parameters α^k and Γ^k emerge from elementary calculations that we now outline.

Let the mean service time at station k be $1/\mu_k$, and denote its squared coefficient of variation by c_k^S; the corresponding parameters for the exogenous interarrival times are $1/\alpha_k^0$ and c_k^A. A FSAT for the renewal process A_j then takes the form

$$\sup_{0 \leq t \leq T} |A_j(t) - \alpha_j^0 t - \alpha_j^0 c_j^A W_j^0(t)| = o(T^{1/r}), \quad \text{as } T \uparrow \infty,$$

and the independence of A_j, $j = 1, \ldots, K$, yields (3.3) for $k = 0$, with means $\alpha^0 = (a_k^0)$ and covariance matrix

$$\Gamma_{j\ell}^0 = \alpha_j^0 c_j^A \delta_{j\ell}, \quad j, \ell = 1, \ldots, K \tag{3.9}$$

Similar FSAT's apply to each S_k, with asymptotic mean μ_k and variance $\mu_k c_k^S$. They are combined with FSAT's for sums of i.i.d. vectors, as in (3.6), to FSAT's for compound renewal processes, as in (3.8). The asymptotic means and covariances are calculated as follows.

Let $r^k(\ell)$, $l = 1, 2, \ldots$, denote the summands of R^k in (3.6). By Wald's identities,

$$\begin{aligned} E[F^k(t)] &= E[S_k(t)]\, E[r^k(1)], \\ \mathrm{Cov}[F^k(t)] &= E[S_k(t)]\, \mathrm{Cov}[r^k(1)] + \mathrm{Var}[S_k(t)]\, E[r^k(1)] E[r^k(1)]'. \end{aligned}$$

Based on the multinomial distribution of the $\xi^k(\ell)$'s,

$$\begin{aligned} E[r_j^k(1)] &= p_{kj} - \delta_{kj}, \\ \mathrm{Cov}[r_j^k(1), r_\ell^k(1)] &= \mathrm{Cov}[\xi_j^k(1), \xi_\ell^k(1)] = p_{kj}\,(\delta_{j\ell} - p_{k\ell})\,. \end{aligned}$$

Finally, the asymptotic covariance of F^k, $\Gamma^k = [\Gamma_{j\ell}^k]$, comes out to be

$$\Gamma_{j\ell}^k = \mu_k p_{kj}\,(\delta_{j\ell} - p_{k\ell}) + \mu_k c_k^S\,(p_{kj} - \delta_{kj})\,(p_{k\ell} - \delta_{k\ell})\,, \; j, \ell = 1, \ldots K, \tag{3.10}$$

for $k = 1, \ldots K$.

3.3 Preliminaries

We briefly recall some concepts and results (from Part I), which are used in the sequel.

3.3.1 Traffic Equations and Bottlenecks

The *effective* arrival rate vector λ is the unique solution to the (nonlinear) traffic equations

$$\lambda = \alpha^0 + P'(\lambda \wedge \mu) \tag{3.11}$$

Its kth coordinate, λ_k, represents the long-run average arrival rate to station k (see Section 2.4 of Part I). The quantity $\rho_k = \lambda_k/\mu_k$ is called the *traffic intensity* at station j. Station k is a *nonbottleneck* if $\rho_k < 1$, *balanced bottleneck* if $\rho_k = 1$, and *strict bottleneck* if $\rho_k > 1$. Denote by $a = \{k : \rho_k < 1\}$, $b = \{k : \rho_k = 1\}$, and $c = \{k : \rho_k > 1\}$, the set of nonbottlenecks, ballanced bottlenecks and strict bottlenecks respectively.

3.3.2 The Oblique Reflection Mapping

Let D^K be the set of K-dimensional RCLL (right-continuous and with left limits) functions, and let $D_0^K = \{x \in D^K : x(0) \geq 0\}$. Let P be a $K \times K$ nonnegative matrix with spectral radius strictly less than unity. The oblique reflection mapping (Section 2.2 of Part I) is characterized in terms of

Theorem 3.3.1 For any $X \in \mathcal{D}_0^K$ there exists a unique pair of $(Y, Z) \in \mathcal{D}_0^{2K}$ satisfying at all $t \geq 0$

$$Z(t) = X(t) + [I - P']Y(t) \geq 0, \tag{3.12}$$
$$dY(t) \geq 0, \qquad Y(0) = 0, \qquad \text{and} \tag{3.13}$$
$$\int_0^\infty Z_k(t)\, dY_k(t) = 0, \quad k = 1, \ldots, K \tag{3.14}$$

Introduce the mappings $Y = \Psi_P(X)$ and $Z = \Phi_P(X)$. Then Ψ_P and Φ_P are both Lipschitz continuous on $\mathcal{D}_0^K$ (with respect to the uniform norm on compact subsets of $[0, \infty)$). Furthermore, $Y = \Psi_P(X)$ is the least among the Y's that satisfy (3.12) and (3.13).

Remarks:

(1) $Y = \Psi_P(X)$ also has a fixed-point characterization. It uniquely satisfies

$$Y(t) = \sup_{0 \leq s \leq t} [P'Y(s) - X(s)]^+, \quad t \geq 0.$$

(See Harrison and Reiman (1981).)

(2) When the dependence on P is obvious, we write Ψ and Φ instead of Ψ_P and Φ_P.

3.3.3 *Reflected Brownian Motion on the Orthant*

Let X be a K-dimensional Brownian motion, starting at $x = X(0) \geq 0$, with drift vector θ and covariance matrix Γ. The process $Z = \{Z(t), t \geq 0\}$, whose sample paths are determined by $Z = \Phi_P(X)$, is known as a Reflected Brownian motion on the nonnegative K-dimensional orthant. It will be denoted by $Z = RBM_x(\theta, \Gamma)$. The process $Y = \{Y(t), t \geq 0\}$, with sample paths $Y = \Psi_P(X)$, is called the regulator of $RBM_x(\theta, \Gamma)$.

The process $RBM_x(\theta, \Gamma)$, $x \geq 0$, is a diffusion process (strong Markov process with continuous sample paths). The kth component Y_k of its regulator Y is its local time on the orthant's face $\{z \geq 0 : z'e^k = 0\}$.

Remarks:

(1) We believe that RBM is Harris recurrent if and only if $[I - P']^{-1}\theta \leq 0$. It has been proved that RBM is positive recurrent if and only if the inequalities are all strict, in which case RBM enjoys a unique stationary/limiting distribution with a density.

(2) Suppose that $[I - P']^{-1}\theta < 0$ and that $P_{kk} = 0$, $k = 1, ..., K$. Under the structural constraints

$$2\Gamma_{jk} = -(P_{kj}\Gamma_{kk} + P_{jk}\Gamma_{jj}) \qquad \text{for} \qquad j \neq k, \tag{3.15}$$

the stationary density has the product-form

$$f(z) = \prod_{k=1}^{K} \eta_k e^{-\eta_k z_k}, \qquad z = (z_1, ..., z_K) \geq 0, \tag{3.16}$$

where $\eta = (\eta_k)$ is given by

$$\eta = -diag(\Gamma)^{-1}[I - P']^{-1}\theta \tag{3.17}$$

Thus, at stationarity or in the limit, Z with (3.15) has independent coordinates, the kth one being exponential with mean $1/\eta_k$. (The constraints (3.15) are necessary and sufficient for the product form (3.16).)

3.4 The Main Results

We start with FSAT in Section 3.4.1 and FLIL in Section 3.4.2. These imply fluid approximations in Section 3.4.3 and diffusion approximations Section 3.4.4.

3.4.1 *Functional Strong Approximations*

Theorem 3.4.1 Suppose that the primitives satisfy the FSAT (3.3) with $r > 2$. Then

$$\sup_{0 \leq t \leq T} |Z(t) - \tilde{Z}(t)| = o(T^{1/r'}), \quad \text{as } T \uparrow \infty, \tag{3.18}$$

where Z is the queue length process in (3.1)-(3.2), $\tilde{Z} = RBM_{Z(0)}(\theta, \Gamma)$ given by

$$\tilde{Z} = \tilde{X} + [I - P']\tilde{Y}, \tag{3.19}$$

$$\tilde{X}(t) = Z(0) + \theta t + \Gamma^{1/2} W(t), \quad t \geq 0, \tag{3.20}$$

$$\theta = \alpha^0 + [P' - I]\mu, \tag{3.21}$$

$$\Gamma = \Gamma^0 + \sum_{k=1}^{K} (\rho_k \wedge 1)\Gamma^k, \tag{3.22}$$

$$\tilde{Y} = \Phi_P(\tilde{X}), \tag{3.23}$$

and W is a K-dimensional standard Wiener process. In (3.18) we can choose $r' = r$ if $r < 4$ and any $r' < 4$ when $r \geq 4$.

Remarks:

(1) For $r \geq 4$, the bound of the FSAT (3.18) can be improved as follows:

$$\sup_{0 \leq t \leq T} |Z(t) - \tilde{Z}(t)| = O((T\text{loglog}T)^{1/4}(\log T)^{1/2}) \quad \text{as } T \uparrow \infty.$$

[This can be justified by taking $g(T) = 5(\log T)^{1/2}$ in (3.54) and (3.55), in the proof of the theorem.] A stronger result would be that (3.18) holds with $r' = r$, $r > 2$, (regardless of whether $r < 4$ or $r \geq 4$). We are not sure, however, whether such a result actually holds.

(2) Let $I(t) = et - B(t)$ represent cumulative idle time, where e is a vector of ones. Then

$$\sup_{0 \leq t \leq T} |I(t) - \text{diag}(\mu)^{-1}\tilde{Y}(t)| = o(T^{1/r'}).$$

(3) The theorem is specialized in Section 3.5 to nonparametric Jackson networks, with Γ expressed in terms of arrival and service rates, transition probabilities and various coefficients of variations.

(4) Consider a queueing network without bottlenecks, namely $b = c = \emptyset$. This is equivalent to $\rho < e$, in which case

$$\rho = \text{diag}(\mu)^{-1}[I - P']^{-1}\alpha^0.$$

Now $\rho < e$ if and only if $[I - P']^{-1}\theta < 0$, for θ in (3.21). Thus, $\tilde{Z}$ has a unique stationary/limit density. Under (3.15), this density has the product form (3.6) with

$$\eta_k = \frac{2\mu_k(1-\rho_k)}{\Gamma_{kk}}, \qquad k = 1, ..., K.$$

(5) We believe that, in general, $\tilde{Z}_a$ is positive recurrent, $\tilde{Z}_b$ is null recurrent and $\tilde{Z}_c$ is transient.

3.4.2 Functional Laws of the Iterated Logarithm

Theorem 3.4.2 Suppose that FLIL (3.5) is satisfied. Then, as $T \uparrow \infty$,

$$\sup_{0\le t\le T} |Z(t) - \bar{Z}(t)| = O(\sqrt{T \log\log T}), \tag{3.24}$$

$$\sup_{0\le t\le T} |B(t) - (\rho \wedge e)t| = O(\sqrt{T \log\log T}), \tag{3.25}$$

where Z and B are, respectively, the queue and busy time processes in (3.1) and (3.2),

$$\bar{Z} = \bar{X} + [I - P']\bar{Y}, \tag{3.26}$$

$$\bar{X}(t) = Z(0) + \theta t, \tag{3.27}$$

$$\bar{Y} = \Phi(\bar{X}), \tag{3.28}$$

and θ is as in (3.21).

Remarks:

(1) The deterministic processes $\bar{Z}_k$, $k = 1, \ldots, K$, represent buffer contents of the linear fluid network (α, P, μ), with initial inventory level $Z(0)$, as introduced in Section 2.4 of Part I, and described in Theorem 2.4.4 there.

(2) The approximations (3.24)-(3.25) still prevail with $\bar{Z}$ being buffer contents of the same linear fluid network (α, P, μ), but with $Z(0) = 0$. [cf. Lemma 2.4.5.] Then $\bar{Z}(t) = [\lambda - \mu]^+ t$, where λ is the effective arrival rate (see Remark 2 that follows Theorem 2.4.4 in Part I).

3.4.3 FSLLN's and Fluid Approximations

Recall the "bar" notation

$$\bar{Z}^n(t) = \frac{1}{n}Z(nt) \qquad \text{and} \qquad \bar{B}^n(t) = \frac{1}{n}B(nt).$$

The following FSLLN is an immediate consequence of Theorem 3.4.2.

Corollary 3.4.3 (FSLLN) Under the assumptions of Theorem 3.4.2,

$$(\bar{Z}^n, \bar{B}^n) \to (\bar{Z}, \bar{B}), \quad \text{u.o.c.}, \quad \text{as } n \to \infty, \tag{3.29}$$

where

$$\bar{B}(t) = (\rho \wedge e)\, t, \qquad \text{and}$$
$$\bar{Z}(t) = [\lambda - \mu]^+ t.$$

Remark: This corollary is, in fact, a consequence of Theorem 2.5.2 from Part I, specialized to a sequence of networks that does not vary with n. Consequently, the characterization of the fluid model in Section 3.4.2 of Part I applies here as well.

3.4.4 FCLT's and Diffusion Approximations

Recall that a, b and c represent nonbottleneck, balanced and strict bottleneck stations, respectively. Introduce the "hat" notation

$$\hat{Z}^n(t) = \sqrt{n}\left[\tfrac{1}{n}Z(nt) - (\lambda - \mu)^+ nt\right],$$
$$\hat{B}^n(t) = \sqrt{n}\left[(\rho \wedge e)t - \tfrac{1}{n}B(nt)\right].$$

The following FCLT is a consequence of Theorem 3.4.1.

Corollary 3.4.4 (FCLT) Under the assumption of Theorem 3.4.1, we have

$$(\hat{Z}^n, \hat{B}^n) \xrightarrow{d} (\hat{Z}, \hat{B}), \quad \text{as } n \to \infty, \tag{3.30}$$

where

$$\hat{Z}_a = 0, \tag{3.31}$$

$$\hat{Z}_b = \hat{X}_b(t) + [I - \hat{P}'_b]\hat{Y}_b(t), \tag{3.32}$$

$$\hat{X}_b = \xi_b + P'_{ab}[I - P'_a]\xi_a, \tag{3.33}$$

$$\hat{X}_b = \xi(t) = \Gamma^{1/2}W(t), \tag{3.34}$$

$$\hat{Y}_b = \Psi_{\hat{P}_b}(\hat{X}_b), \tag{3.35}$$

$$\hat{P}_b = P_b + P_{ba}[I - P_a]^{-1}P_{ab}, \tag{3.36}$$

$$\hat{Z}_c = \xi_c + P'_{ac}[I - P'_a]^{-1}\xi_a - \hat{P}'_{bc}\hat{Y}_b, \tag{3.37}$$

$$\hat{P}_{bc} = P_{bc} + P_{ba}[I - P_a]^{-1}P_{ac}, \tag{3.38}$$

$$\hat{B}_a = \text{diag}(\mu_a)^{-1}[I - P'_a]^{-1}[\xi_a - P'_{ba}\hat{Y}_b], \tag{3.39}$$

$$\hat{B}_b = -\text{diag}(\mu_b)^{-1} \cdot \hat{Y}_b, \tag{3.40}$$

$$\hat{B}_c = 0, \tag{3.41}$$

Γ is the covariance matrix in (3.22), and W is a K-dimensional standard Wiener process as in (3.20)).

Remarks:

(1) The corollary demonstrates that the diffusion limit of queue length vanishes at the nonbottlenecks a. The diffusion approximation of the balanced subnetwork b is a $|b|$-dimensional reflected Brownian motion. The diffusion limit for queue lengths at strict bottlenecks requires centering. This is because Z_c builds up at a rate of $(\lambda_c - \mu_c)$, which is also the buildup rate of the corresponding fluid approximation [cf. Corollary 3.4.3]. After both centering and rescaling, Z_c^n converges to the semimartingale Z_c in (3.37): the martingale component of Z_c is a Brownian motion, which is associated with a and c; its bounded-variation component is nonincreasing and is associated with b [see (3.35)].

(2) The corollary provides limits in light-traffic (no bottlenecks: $b = c = \emptyset$). For example, the centered and rescaled busy time processes $\hat{B}^n$ converges weakly to the driftless Brownian motion

$$\hat{B} = \text{diag}(\mu)^{-1}[I - P']^{-1}\xi,$$

with ξ in (3.34).

(3) Proposition 2.4.2 in Part I guarantees the existence of the inverse $[I - P_a]^{-1}$.

3.5 Fitting Parametes

In Section 3.5.1 we provide two approximations to the nonparametric Jackson queueing network from Section 3.2.3: a first order approximation by a fluid network, which is supported by the FLIL Theorem 3.4.2, and a refined second order approximation by an RBM, which is based on the FSAT Theorem 3.4.1. These are specialized in Section 3.5.2 to the product-form and single-station cases.

3.5.1 Nonparametric Jackson Networks

At a first order, the queue length process Z can be approximated by the buffer content process $\bar{Z}$ of the linear fluid model

$$\begin{aligned} \bar{Z} &= \bar{X} + [I - P']\bar{Y}, \\ &\quad \bar{X}(t) = Z(0) + [\alpha^0 + (P' - I)\mu]\, t, \\ &\quad \bar{Y} = \Phi(\bar{X}). \end{aligned}$$

The order of the approximation is given by

$$\sup_{0 \le t \le T} |Z(t) - \bar{Z}(t)| = O(\sqrt{T \log\log T}), \quad \text{as } T \uparrow \infty.$$

A refined second-order approximation is the RBM

$$\begin{aligned} \tilde{Z} &= \tilde{X} + [I - P']\tilde{Y}, \\ &\quad \tilde{X}(t) = Z(0) + [\alpha^0 + (P' - I)\mu]\, t + \Gamma^{1/2}\, W(t), \\ &\quad \tilde{Y} = \Phi(\tilde{X}), \end{aligned}$$

where W is a K-dimensional Wiener process. The covariance matrix $\Gamma = [\Gamma_{k\ell}]$ is calculated from (3.9), (3.10) and (3.22), and takes the form

$$\Gamma_{k\ell} = \alpha_k^0 c_k^A \delta_{k\ell} + \sum_{j=1}^{K} (\lambda_j \wedge \mu_j) \left[p_{jk}(\delta_{k\ell} - p_{j\ell}) + c_j^S (p_{jk} - \delta_{jk})(p_{j\ell} - \delta_{j\ell}) \right].$$

Assuming finite moments of order $r > 2$, and an ambient probability space on which all our stochastic elements are defined, the error is quantified by the pathwise bounds

$$\sup_{0 \le t \le T} |Z(t) - \tilde{Z}(t)| = o(T^{1/r}), \quad \text{as } T \uparrow \infty.$$

3.5.2 Product Form and Single Station

Consider the case of no-bottlenecks ($\rho < e$). It is then plausible to approximate the steady-state of Z by that of $\tilde{Z}$. In particular, when (3.15) approximately applies, and in view of Remark 4 below Theorem 3.4.1, the steady-state queue-lengths $Z_k(\infty)$, $k = 1, \ldots, K$, are approximately independent exponential, with means

$$E[Z_k(\infty)] \approx \frac{\Gamma_{kk}}{2\mu_k(1 - \rho_k)}.$$

The special case of a single station ($K = 1$) reduces to

$$E[Z(\infty)] \approx \frac{\rho}{(1 - \rho)} \left[\frac{(c^A + c^S)}{2} + \frac{p(1 - c^S)}{2} \right] (1 - p),$$

where p is the probability of feedback, and $\rho = \alpha^0/[\mu(1-p)]$. Finally, $p = 0$ reduces to the classical approximation

$$E[Z(\infty)] \approx \frac{\rho}{(1 - \rho)} \frac{(c^A + c^S)}{2},$$

which originated in Kingman's pioneering work on the G/G/1 queue in heavy traffic.

3.6 Proof of the Main Results

We start with the proof of Theorem 3.4.2. We then use it to prove Theorem 3.4.1. Finally we prove Corollary 3.4.3 and Corollary 3.4.4.

Note that Theorem 3.4.1 actually implies Theorem 3.4.2, if the stronger assumption (3.5) is imposed.

Proof of Theorem 3.4.2

First rewrite (3.1) as

$$Z(t) = X(t) + [I - P']Y(t), \tag{3.42}$$

where

$$\begin{aligned} X(t) &= Z(0) + \theta t + [F^0(t) - \alpha^0 t] \\ &\quad + \sum_{j=1}^{k} \big[F^j(B_j(t)) - \alpha^j B_j(t)\big], \qquad (3.43) \\ Y(t) &= \mathrm{diag}(\mu)\big[et - B(t)\big], \qquad (3.44) \end{aligned}$$

and θ is as defined in (3.21). Applying the FLIL assumption (3.5) to (3.43) yields

$$||\bar{X} - X||_T = \sup_{0\le t\le T} |X(t) - \bar{X}(t)| = O(\sqrt{T log log T}), \quad T \uparrow \infty, \tag{3.45}$$

where $\bar{X}$ is as defined in (3.27). Thus, (3.45) and the Lipschitz continuity of the reflection mapping (Theorem 3.3.1) lead to equality (3.24) in the theorem, and

$$||Y - \bar{Y}||_T = \sup_{0\le t\le T} |Y(t) - \bar{Y}(t)| = O(\sqrt{T log log T}) \tag{3.46}$$

The proof of equality (3.25) in the theorem is now completed by combining (3.44) and (3.46), with equality (3.47) to be proved in the following lemma.

Lemma 3.6.1 Let $\bar{Y} = \Psi(\bar{X})$ with $\bar{X}$ as in (3.27). Then there exists an $M > 0$ such that

$$\sup_{0\le t<\infty} \|\bar{Y}(t) - (\mu - \lambda)^+ t\| \le M, \tag{3.47}$$

where λ is the effective arrival rate that solves (3.4).

Proof. Let $\bar{X}^*(t) = \theta t$. Then $\Psi(\bar{X}^*)(t) \equiv (\mu - \lambda)^+ t$ (see the Remark below Theorem 2.4.4 in Part I). By the Lipschitz property of the reflection mapping, there exists an $M' > 0$ such that

$$\|\Psi(\bar{X}) - \Psi(\bar{X}^*)\| \le M' \|\bar{X} - \bar{X}^*\| = M' \|Z(0)\| \equiv M.$$

Proof of Theorem 3.4.1

Proving the theorem for $r < 4$ establishes it for $r \ge 4$ since (2.23) for $r \ge 4$ implies it for all $r < 4$. We thus restrict attention to $r < 4$.

First force (3.1) into the form (3.12) via

$$Z(t) = X(t) + [I - P']Y(t),$$

where

$$\begin{aligned} X(t) &= Z(0) + \big[\alpha^0 + (P' - I)\mu\big]t + \big[F^0(t) - \alpha^0 t - (\Gamma^0)^{1/2} W^0(t)\big] \\ &\quad + \sum_{k=1}^{K} \big[F^j(B_k(t)) - \alpha^j B_k(t) - (\Gamma^j)^{1/2} W^k(B_k(t))\big] \\ &\quad + \sum_{k=1}^{K} (\Gamma^k)^{1/2} \big[W^k(B_k(t)) - W^k\big((\rho_k \wedge 1)t\big)\big] \\ &\quad + (\Gamma^0)^{1/2} W^0(t) + \sum_{k=1}^{K} (\Gamma^k)^{1/2} W^k\big((\rho_k \wedge 1)t\big), \qquad (3.48) \\ Y(t) &= \mathrm{diag}(\mu)\big[et - B(t)\big]. \qquad (3.49) \end{aligned}$$

Let

$$\begin{aligned} \widetilde{X}(t) &= Z(0) + \theta t + (\Gamma^0)^{1/2} W^0(t) \\ &\quad + \sum_{k=1}^{K} (\Gamma^k)^{1/2} W^k\big((\rho_k \wedge 1)t\big), \qquad (3.50) \\ &\equiv Z(0) + \theta t + \Gamma^{1/2} W(t). \qquad (3.51) \end{aligned}$$

with θ as in (3.21), and (3.51) being a defining relation for the standard Brownian motion W. Once we prove that

$$\sup_{0\le t\le T} |X(t) - \widetilde{X}(t)| = o(T^{1/r}), \tag{3.52}$$

then we immediately deduce (3.18) from the Lipschitz continuity of the mapping Ψ. By the FSAT assumption (2.23), to prove (3.52) it suffices to show (note that $|B(t) - B(s)| \le |t - s|$) that

$$\sup_{0\le t\le T} |W_k^j(B_j(t)) - W_k^j((\rho_j \wedge 1)t)| = o(T^{1/r}),$$

for all $j, k = 1, \ldots K$. This we do now by showing that, for any Wiener process W on our ambient probability space,

$$\sup_{0\le t\le T} |W(B_j(t)) - W((\rho_j \wedge 1)t)| = o(T^{1/r}) \tag{3.53}$$

for all $j = 1, 2, \ldots, K$. We start with a lemma, whose proof is postponed to the end of this subsection.

Lemma 3.6.2 Let $W = \{W(t), t \geq 0\}$ be a Wiener process. Then for any $\delta > 0$ there exists a constant $C = C(\delta) > 0$ such that the inequality

$$P\left\{\sup_{0\leq u,v\leq T;\ |u-v|\leq h} |W(u) - W(v)| \geq x\sqrt{h}\right\} \leq C(1 + \frac{T}{h})e^{-x^2/(2+\delta)}$$

holds for every $T > 0$ and $0 < h < T$.

Taking $h = h(T)$ and $X = g(T)$ in the above inequality and then applying the Borel-Cantelli Lemma yields,

Lemma 3.6.3 Let $W = \{W(t), t \geq 0\}$ be is a Wiener process. Suppose that there exists a pair of functions h and g satisfying

$$\frac{g(T)\sqrt{h(T)}}{T^{1/r}} \to 0, \qquad \text{as } T \to \infty, \tag{3.54}$$

$$\sum_{T=1}^{\infty} \frac{T}{h(T)} e^{-[g(T)]^2/(2+\delta)} < \infty, \qquad \text{as } T \to \infty. \tag{3.55}$$

Then with probability one,

$$\sup_{0\leq u,v\leq T;|u-v|\leq h(T)} |W(u) - W(v)| = o(T^{1/r}), \qquad \text{as } T \uparrow \infty. \tag{3.56}$$

When $2 < r < 4$,

$$h(T) = M\sqrt{TloglogT} \qquad \text{and} \qquad g(T) = T^{\frac{1}{2}(\frac{1}{r}-\frac{1}{4})}$$

clearly satisfy (3.54) and (3.55). On the other hand, from equality (3.25) of Theorem 3.4.2, we know that there exists an $M > 0$ such that

$$\sup_{0\leq t\leq T} |B_j(t) - (\rho_j \wedge 1)t| \leq M\sqrt{TloglogT}.$$

Now applying Lemma 3.6.3 yields (3.53).

Finally, if we note that W^j, $j = 0, 1, ..., K$, are independent Wiener processes, then (3.50) implies that the covariance matrix Γ is of form (3.22).

Proof of Lemma 3.6.2 Noting that

$$\{(u, v) : |u - v| \leq h, \quad 0 \leq u \leq T, \quad 0 \leq v \leq T\}$$

is a subset of

$$\{(u, v) : 0 \leq u \leq T, 0 \leq v - u \leq h\} \bigcup \{(u, v) : 0 \leq v \leq T, 0 \leq u - v \leq h\},$$

we can use Lemma 1.2.1 of Csörgö and Revesz [1981] with T replaced by $T + h$ and C replaced by $C/2$ to obtain the desired results. □

Proof of Corollary 3.4.3

We take $\bar{Z}$ as defined in Corollary 3.4.3. First, it follows from (3.24) and (3.25) of Theorem 3.4.2 (also note Remark (1) below the theorem) that for any fixed $T > 0$,

$$\sup_{0\leq t\leq T} \left|\frac{1}{n}Z(nt) - \frac{1}{n}\bar{Z}(nt)\right| = O\left(\sqrt{\frac{1}{n}\log\log n}\right), \tag{3.57}$$

$$\sup_{0\leq t\leq T} \left|\frac{1}{n}B(nt) - (\rho \wedge e)\, t\right| = O\left(\sqrt{\frac{1}{n}\log\log n}\right). \tag{3.58}$$

These two equalities clearly imply the convergence of (3.29).

Proof of Corollary 3.4.4

For fixed $T > 0$, the strong approximation (3.18) implies that

$$\sup_{0\leq t\leq T} \left|\frac{1}{\sqrt{n}}Z(nt) - \frac{1}{\sqrt{n}}\tilde{Z}(nt)\right| = \frac{o(1)}{n^{(r-2)/(2r)}}, \quad \text{as } n \uparrow \infty.$$

Thus, the proof for Corollary 3.4.4 amounts to proving a central limit theorem for the reflected Brownian motion $\tilde{Z}$ as defined through (3.19)-(3.23), which will be stated as Proposition 3.6.4 below.

To simplify notations, we remove the tilde notation from the reflected Brownian motion; namely, the reflected Brownian motion will be represented by

$$Z = X + [I - P']Y, \tag{3.59}$$

$$X(t) = Z(0) + \theta\, t + \Gamma\, W(t), \tag{3.60}$$

$$\theta = \alpha + [P' - I]\mu, \tag{3.61}$$

$$Y = \Phi(X). \tag{3.62}$$

We define

$$B(t) = e\, t - \text{diag}(\mu^{-1})Y(t), \tag{3.63}$$

and "hat" notations:

$$\hat{Z}^n(t) = \frac{1}{\sqrt{n}}\left[Z(nt) - (\lambda - \mu)^+ nt\right],$$

$$\hat{B}^n(t) = \frac{1}{\sqrt{n}}\left[(\rho \wedge e)t - B(nt)\right], \qquad \text{and}$$

$$\hat{Y}^n(t) = \frac{1}{\sqrt{n}}Y(nt).$$

Proposition 3.6.4 For the processes defined in (3.59)-(3.62), the convergence (3.30) prevails with the limiting processes defined by (3.31)-(3.40).

First we state and prove two lemmas.

Lemma 3.6.5 Recall that a is the set of nonbottleneck stations. Then we have

$$\hat{Z}_a^n \xrightarrow{d} 0, \quad \text{as } n \to \infty \tag{3.64}$$

Proof. The proof proceeds as follows. First, we prove the convergence (3.64) for the set of sub-critical stations (those for which $\theta_j < 0$; see Section 2.4 of the previous chapter). Then we consider a subnetwork which is obtained from the original network by removing all sub-critical stations, and then we prove the convergence (3.64) for those stations that are sub-critical stations in the subnetwork. We continue with this process unless we reach a subnetwork which does not have sub-critical stations, and then, by Propositions 2.4.2 and 2.4.3 in the previous chapter, we know that the lemma is proved. This process is very similar to that used in the proof of Theorem 2.4.4 in last chapter. So we will only provide the first step here, i.e., to prove the convergence of (3.64) for the set of sub-critical stations.

Consider a sub-critical station j, i.e., $\theta_j < 0$. Let $\epsilon_n = \max\{\bar{Z}_k^n(0); k = 1, ..., K\}$ and note that $\epsilon_n \to 0$. Then introduce for $t \geq 0$,

$$\nu_j^n(t) = \sup\{s \leq t : \bar{Z}_j^n(s) \leq \epsilon_n\}$$

(abbreviated as ν_j^n when convenient and well defined because the set over which the supremum is taken always contains $s = 0$: $\hat{Z}_j^n(0) \leq \epsilon_n$). Now $\hat{Z}_j^n$ is a continuous function. It follows from the definition of $\nu_j^n(t)$ that $\hat{Z}_j^n\left(\nu_j^n\right) = \epsilon_n$. Furthermore, if $\nu_j^n(t) < t$ then $\hat{Z}_j^n(s) > \epsilon_n \geq 0$ for $s \in [\nu_j^n(t), t]$. Using the complementarity condition (3.14), applied to the "hat" representation of (3.59), implies that $\hat{Y}_j^n\left(\nu_j^n(t)\right) = \hat{Y}_j^n(t)$, which also holds when $\nu_j^n(t) = t$. One utilizes all this in

$$\begin{aligned} -\epsilon_n &\leq \hat{Z}_j^n(t) - \hat{Z}_j^n\left(\nu_j^n\right) \\ &= \hat{\xi}_j^n(t) - \hat{\xi}_j^n\left(\nu_j^n\right) + n\theta_j(t - \nu_j^n) - \sum_{k=1}^{J} p_{kj}\left[\hat{Y}_k^n(t) - \hat{Y}_k^n\left(\nu_j^n\right)\right] \\ &\leq \hat{\xi}_j^n(t) - \hat{\xi}_j^n\left(\nu_j^n\right) + n\theta_j(t - \nu_j^n), \end{aligned} \tag{3.65}$$

where

$$\hat{\xi}^n(t) = \Gamma\frac{1}{\sqrt{n}}W(nt) \tag{3.66}$$

Consequently, for $t \geq 0$,

$$0 \leq (-\theta_j)[t - \nu_j^n(t)] \leq \frac{\epsilon_n}{n} + \frac{1}{n}\hat{\xi}_j^n(t) - \frac{1}{n}\hat{\xi}_j^n\left(\nu_j^n\right) \tag{3.67}$$

The functional law of iterated logarithm for Brownian motion guarantees that both $\hat{\xi}_j^n(t)/n$ and $\hat{\xi}_j^n(\nu_j^n(t))/n$ converge uniformly to zero on any compact subset of $[0, \infty)$. Thus, one concludes from (3.67) and $\theta_j < 0$ that

$$\nu_j^n(t) \to t \quad u.o.c. \tag{3.68}$$

Clearly,

$$\hat{\xi}^n \xrightarrow{d} \hat{\xi}, \quad \text{as } n \to \infty\,. \tag{3.69}$$

Going back to (3.65), we have

$$0 \leq \hat{Z}_j^n(t) \leq \hat{\xi}_j^n(t) - \hat{\xi}_j^n\left(\nu_j^n\right) + \hat{Z}_j^n\left(\nu_j^n\right),$$

for all $t \geq 0$. Finally, $0 \leq \hat{Z}_j^n\left(\nu_j^n\right) \leq \epsilon_n$, combined with (3.68) and (3.69), establishes the convergence (3.64) for all sub-critical stations. □

Lemma 3.6.6 Recall that c is the set of strict bottlenecks. Then we have

$$\hat{Y}_c^n \xrightarrow{d} \hat{\xi}, \quad \text{as } n \to \infty \tag{3.70}$$

Proof. Let β be the union of b and c, i.e., β is the set of all bottleneck stations. The lemma will be proved in two steps. First, we prove that there exists a sequence of processes $\tilde{Y}_\beta^n$, $n = 1, 2, ...$, in $\mathcal{C}^{|\beta|}$ that dominates $\hat{Y}_\beta^n$, $n = 1, 2, ...$, and converges weakly to a finite process. Second, we show the desired convergence (3.70).

To the end of proving the first step, we rewrite $\hat{Z}^n$ of (3.59) in a and β block form,

$$\begin{aligned} \hat{Z}_a^n(t) &= \hat{Z}_a^n(0) + \theta_a\sqrt{n}t + \hat{\xi}_a^n(t) \\ &\quad - P'_{\beta a}\hat{Y}_\beta^n(t) + [I - P'_a]\hat{Y}_a^n(t), \end{aligned} \tag{3.71}$$

$$\begin{aligned} \hat{Z}_\beta^n(t) + (\lambda_\beta - \mu_\beta)\sqrt{n}t &= \hat{Z}_\beta^n(0) + \theta_\beta\sqrt{n}t + \hat{\xi}_\beta^n(t) \\ &\quad - P'_{a\beta}\hat{Y}_a^n(t) + [I - P'_\beta]\hat{Y}_\beta^n(t), \end{aligned} \tag{3.72}$$

where $\hat{\xi}^n$ is defined in (3.66).

Since $\sigma(P) < 1$, the inverse of $(I - P'_a)$ exists [Proposition 2.4.2 of the last chapter]. Solving for $\hat{Y}_a^n$ in (3.71) and substituting the outcome into (3.72) yields

$$\hat{Z}_\beta^n(t) + (\lambda_\beta - \mu_\beta)\sqrt{n}t = \chi_\beta^n(t) + (\lambda_\beta - \mu_\beta)\sqrt{n}t + [I - \hat{P}'_\beta]\hat{Y}_\beta^n, \tag{3.73}$$

where

$$\begin{aligned} \chi_\beta^n(t) &= [\hat{Z}_\beta^n(0) + P'_{a\beta}(I - P'_a)^{-1}\hat{Z}_a^n(0)] - P'_{a\beta}(I - P'_a)^{-1}\hat{Z}_a^n \\ &\quad + [\hat{\xi}_\beta^n + P'_{a\beta}(I - P'_a)^{-1}\hat{\xi}_a^n], \\ \hat{P}_\beta &= P_\beta + P_{\beta a}(I - P_a)^{-1}P_{a\beta}, \end{aligned} \tag{3.74}$$

and we used the fact that

$$\theta_\beta + P'_{a\beta}(I - P'_a)^{-1}\theta_a = \lambda_\beta - \mu_\beta \tag{3.75}$$

By the definition of β, we know that $\lambda_\beta - \mu_\beta \geq 0$. Therefore, by the least element characterization of mapping Ψ (Theorem 3.3.1), we have

$$\hat{Y}^n_\beta = \Psi_{\hat{P}_\beta}(\chi^n_\beta + (\lambda_\beta - \mu_\beta)\sqrt{n}e) \leq \tilde{Y}^n_\beta \equiv \Psi_{\hat{P}_\beta}(\chi^n_\beta). \tag{3.76}$$

By (3.69) and Lemma 3.6.5, we can see that χ^n_β converges weakly to a Brownian motion, and therefore, the continuous mapping theorem implies that $\tilde{Y}^n_\beta$ converges weakly to a finite limit (in fact, the regulator of a reflected Brownian motion). Now, we have completed the first step.

To complete the proof, we rewrite $\hat{Z}^n$ of (3.59) in blocks a, b and c:

$$\begin{aligned}
\hat{Z}^n_a(t) &= \hat{Z}^n_a(0) + \theta_a\sqrt{n}t + \hat{\xi}^n_a(t) \\
&\quad - P'_{ba}\hat{Y}^n_b(t) - P'_{ca}\hat{Y}^n_c(t) + [I - P'_a]\hat{Y}^n_a(t), \quad (3.77)\\
\hat{Z}^n_b(t) &= \hat{Z}^n_b(0) + \theta_b\sqrt{n}t + \hat{\xi}^n_b(t) \\
&\quad - P'_{ab}\hat{Y}^n_a(t) - P'_{cb}\hat{Y}^n_c(t) + [I - P'_b]\hat{Y}^n_b(t), \quad (3.78)\\
\hat{Z}^n_c(t) + (\lambda_c - \mu_c)\sqrt{n}t &= \hat{Z}^n_c(0) + \theta_c\sqrt{n}t + \hat{\xi}^n_c(t) \\
&\quad - P'_{ac}\hat{Y}^n_a(t) - P'_{bc}\hat{Y}^n_b(t) + [I - P'_c]\hat{Y}^n_c(t) \quad (3.79)
\end{aligned}$$

Then we solve for $\hat{Y}^n_a$ in (3.77), and substitute the outcome into (3.79) to obtain

$$\begin{aligned}
\hat{Z}^n_c(t) + (\lambda_c - \mu_c)\sqrt{n} &= \hat{X}^n(t) - \hat{P}'_{bc}\hat{Y}^n_b + (\lambda_c - \mu_c)\sqrt{n}t \\
&\quad + [I - \hat{P}'_c]\hat{Y}^n_c,
\end{aligned} \tag{3.80}$$

where

$$\begin{aligned}
\hat{X}^n_c(t) &= [\hat{Z}^n_c(0) + P'_{ac}(I - P'_a)^{-1}\hat{Z}^n_a(0)] - P'_{ac}(I - P'_a)^{-1}\hat{Z}^n_a \\
&\quad + [\hat{\xi}^n_c + P'_{ac}(I - P'_a)^{-1}\hat{\xi}^n_a], \\
\hat{P}_c &= P_c + P_{ca}(I - P_a)^{-1}P_{ac},
\end{aligned}$$

$\hat{P}_{bc}$ is defined in (3.38), and we also used (3.75).

Applying Remark (1) that succeeds Theorem 3.3.1 to (3.80) and observing inequality (3.76) yields

$$\begin{aligned}
0 \leq \hat{Y}^n_c(t) &= \sup_{0\leq s\leq t}\left[\hat{P}'_c\hat{Y}^n_c(s) - \hat{X}^n(s) + \hat{P}'_{bc}\hat{Y}^n_b(s) - (\lambda - \mu)\sqrt{n}s\right]^+ \\
&\leq \sup_{0\leq s\leq t}\left[\hat{P}'_c\tilde{Y}^n_c(s) - \hat{X}^n_c(s) + \hat{P}'_{bc}\tilde{Y}^n_b(s) - (\lambda_c - \mu_c)\sqrt{n}s\right]^+ .
\end{aligned} \tag{3.81}$$

Now note that all processes $\hat{X}^n_c$, $\tilde{Y}^n_b$ and $\tilde{Y}^n_c$ converge weakly to finite limits, and that $\lambda_c - \mu_c > 0$ (since c is the set of strict bottlenecks). Then applying the functional limit theorem for the supremum (Theorem 6.1 of Whitt [1980]), we prove that the process defined by the right hand side of inequality (3.81) converges weakly to zero, therefore, the convergence (3.70). □

Proof of Proposition 3.6.4 First solving for $\hat{Y}^n_a$ in (3.77) and substituting the outcome into (3.78) yields

$$\hat{Z}^n_b(t) = \hat{X}^n_b(t) + [I - \hat{P}'_b]\hat{Y}^n_b, \tag{3.82}$$

where

$$\begin{aligned}
\hat{X}^n_b(t) &= [\hat{Z}^n_b(0) + P'_{ab}(I - P'_a)^{-1}\hat{Z}^n_a(0)] - P'_{ab}(I - P'_a)^{-1}\hat{Z}^n_a \\
&\quad + \left[\hat{\xi}^n_b + P'_{ab}(I - P'_a)^{-1}\hat{\xi}^n_a\right] - \left[P'_{cb} + P'_{ab}(I - P'_a)P'_{ca}\right]\hat{Y}^n_c,
\end{aligned} \tag{3.83}$$

$\hat{P}_b$ is defined in (3.36), and we used (3.75).

It follows from (3.69) and Lemmas 3.6.5 and 3.6.6 that $\hat{X}^n_b$ in (3.82) converges weakly to the Brownian motion $\hat{X}_b$ in (3.33). The continuous mapping theorem then implies that

$$\left(\hat{Y}^n_b = \Psi_{\hat{P}_b}\left(\hat{X}^n_b\right),\ \hat{Z}^n_b = \Phi_{\hat{P}_b}\left(\hat{X}^n_b\right)\right)$$

converges weakly to the reflected Brownian motion $(\hat{Y}_b, \hat{Z}_b)$ defined through (3.32)-(3.36). Next, the weak convergences of (3.69), $\hat{Y}^n_a$, $\hat{Y}^n_b$, and $\hat{Z}^n_a$, applied to (3.80), imply the weak convergence of $\hat{Z}^n_c$ with the limit $\hat{Z}^n_c$ in (3.37). The convergence of $\hat{B}^n$ is by observing (3.76) and

$$\hat{B}^n(t) = -\mathrm{diag}(\mu^{-1})\left[\hat{Y}^n(t) - (\mu - \lambda)^+\sqrt{n}t\right].$$

Finally, we remark that all of the convergences actually hold jointly.

3.7 References, Possible Extensions and Future Research

General Commentary: We have introduced strong approximations as a unifying framework, at the cost of imposing assumptions that are mathematically too stringent. Indeed, all of our corollaries can be established, individually, under weaker conditions. However, as far as current applications are concerned, such stronger results seem to offer no benefit.

In this chapter we focus on homogeneous single-server open networks. Natural extensions include multi-server, and homogeneous closed and mixed

networks, as well as heterogeneous (multi-class) networks of all kinds. Fluid and diffusion approximations for heterogeneous multi-server nonparametric Jackson networks under heavy traffic are given in Chen and Shanthikumar (1990). Combining their approach with ours easily establishes the strong approximation for the heterogeneous multi-server network, with no heavy traffic assumptions. Fluid and diffusion approximations for nonparametric closed Jackson networks are given in Chen and Mandelbaum [1991b]. As far as FLIL and FSAT for closed networks, FLIL for the queue length process (3.24) still holds, but our approach does not carry over to the FLIL for the busy time process (3.25). Establishing FSAT for closed network takes a different approach; see recent work by Zhang [1993]. This state of affairs is consistent with the fact that the mapping Φ for closed networks is Lipschitz (Dupuis and Ishii [1991]) but Ψ need not be (a counter example appears in Berger and Whitt [1992]). Mixed networks are currently under study by Nguyen [1993]. Heterogeneous open networks present a significant challenge, as apparent from Harrison [1988], Harrison and Nguyen [1990;1992], Dai and Wang [1993] and Whitt [1994].

There is an alternative, deeper but less "user friendly", form of strong approximations, in terms of exponential bounds on the probability of deviations from a central trend. (See Csörgö and Révész [1981] and Glynn [1990]). We have not presented this form here but, with some modification, it can also be handled by our approach.

Work on strong approximations for queueing networks has been rather scarce. We are aware only of Zhang, Hsu and Wang [1990], who analyze a single station with multiple-servers, Zhang [1990], that covers super-critical nonparametric Jackson networks, Horváth [1990], who treats two-station open network, and Glynn and Whitt [1991] that deal with queue in series.

Meyn and Down [1993] proves existence of a stationary distribution for non-parametric closed and open Jackson networks. As in Horváth [1990], one should be able to use strong approximations refinements to show that this stationary distribution, properly normalized, converges to the stationary distribution of a corresponding RBM. (This is actually Kingman's approach in his pioneering work on heavy traffic: he showed convergence to the exponential distribution, which is the stationary distribution of the one-dimensional RBM; the issue is easy to settle for closed networks, as done in Kaspi and Mandelbaum [1989], but it is still open for open networks.

Harrison and Williams [1987] established an integral equation, known as a basic adjoint relation (BAR), to characterize the stationary distribution of the approximating diffusion process for open queueing networks. Harrison, Williams and Chen [1990] extends the result to irreducible closed networks. BAR enables calculations of useful performance measures. On rare occasions, the integral equations can be solved to yield an explicit expression for the stationary distribution (typically of a product-form). Dai and Harrison [1991,1992] have developed computational methods for more general cases.

Diffusion approximations are used not only for performance evaluation, but also for optimal control of queueing networks. Specifically, a queueing control problem is approximated by a diffusion control problem that is easier to handle, and then optimal, or near optimal, solutions to the latter problem are interpreted in the original queueing context. This line of research started in Harrison [1988], and continued for example in Yang [1988], Wein [1990], Harrison and Wein [1990]. Kelly and Laws [1991] expresses the hope that ultimate justification of results that propose asymptotically optimal control will not depend on the Brownian nature of the approximation. It is plausible that the framework of strong approximations is what they are searching for. We also note that Krichagina, Lou, Sethi and Taksar [1991] proved the asymptotic optimality of the diffusion approximation for a failure-prone manufacturing system.

Section 1

There is a voluminous literature on queueing networks, most of which traces back to the seminal works of Jackson [1963], Gorden and Newell [1967], Whittle [1967;1968], BCMP [1975] and Kelly [1979].

Diffusion approximations for queueing systems have been a subject of research for almost 30 years. Readers are referred to Kingman [1965], Iglehart and Whitt [1970a,b], Reiman [1984], Johnson[1983], and Chen and Mandelbaum [1991b]. More references can be found in the papers cited above and in the survey papers by Whitt [1974], Lemoine [1978], Flores [1985] and Glynn [1990].

Consider a non-parametric queueing network whose traffic intensities are all strictly less than unity. The approximation of such a single queueing network commonly entails perturbing (rescaling) it to get an approximating *sequence* of networks, and then taking limits. Such an approximation is informative at stations with traffic intensities that are very close to unity (in heavy traffic); otherwise, the diffusion limit is zero. Our strong approximations, on the other hand, are always applicable: no rescaling is needed, and they give rise to both an approximating diffusion and an error bound. In particular, strong approximations are applicable at stations with traffic intensities significantly less than unity. This is consistent with success of the numerical method developed by Dai and Harrison [1992], which perform well in heavy traffic as well as moderate traffic.

Section 2

The model and the notation are adopted from Harrison and Williams [1987]. The FSAT for renewal processes is Part (ii) of Corollary 3.1, in Csörgö, Horváth and Steinebach [1987]. The FSAT for compound renewal processes is Theorem A in the Appendix to Horváth [1992]. Horváth uses FSAT's for summands of random vectors, which he attributes to Einmal

[1989]. (Note that Horváth [1992] considers the probability-bound form of the FSAT's.)

Section 3

Most of this section is a recapitulation of Sections 2.2 and 2.4 from Part I. Remark (2) in Section 3.3.3 follows from Harrison and Williams [1987].

Section 4

Note that the strong approximation (3.18) implies the functional law-of-iterated-logarithm. The latter result holds under a weaker condition than the former.

Strong approximations can also be developed for a sequence of networks. Then, Corollary 3.4.4 would recover the diffusion limits of Reiman [1984] and Chen and Mandelbaum [1991b].

Section 5

For more details on how to fit a network, and the detailed calculations of the covariance matrix, readers are referred to Harrison [1988] and Harrison and Williams [1987].

Acknowledgments: We would like to thank S.G. Kou for pointing out an error in the earlier proof of Theorem 3.4.1. Hong Chen is supported in part by NSF grant DDM-89-09972, by NSERC grant OG0089698, and by a UBC-HSS Research Grant.

3.8 References

[1] Baskett, F., Chandy, K.M., Muntz, R.R. and Palacois, F.G. (1975). "Open, closed and mixed networks of queues with different classes of customers", *J. Assoc. Comput. Mach.* **22** 248-260.

[2] Berger, A.W., and Whitt, W., (1992). "The Brownian approximation for rate-control throttles and the $G/G/1/C$ queue", *Discrete Event Dynamic Systems: Theory and Applications* **2** 7-60.

[3] Chen, H., and Mandelbaum, A. (1991a). "Discrete flow networks: bottleneck analysis and fluid approximations", *Math of OR* **16** 408-446.

[4] Chen, H., and Mandelbaum, A. (1991b). "Discrete flow networks: diffusion approximations and bottlenecks", *Ann. of Prob.* **19** 1463-1519.

[5] Chen, H., and Shanthikumar, G. (1990). "Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic", *Discrete Event Dynamic Systems: Theory and Applications* (forthcoming).

[6] Csörgö, M., Horváth L. and Steinebach, J. (1987). "Invariance principles for renewal processes", *Ann. Probab.* **15** 1441-1460.

[7] Csörgö, M., and Revesz, P. (1981). *Strong Approximations in Probability and Statistics*, Academic Press, New York, 1981.

[8] Dai, J.G., and Harrison, J.M. (1991). "Steady-state analysis of reflected Brownian motions: characterization, numerical methods and a queueing application", *Annals of Applied Probability* **1** 16-35.

[9] Dai, J.G., and Harrison, J.M. (1991). "Reflected Brownian motion in an orthant: numerical methods and a queueing application", *Annals of Applied Probability* **2** 65-86.

[10] Dai, J.G., and Wang, Y. (1993). "Nonexistence of Brownian models of certain multiclass queueing networks", *Queueing Systems* **13** 41-46.

[11] Dupuis, P., and Ishii, H. (1991). "On when the solution to the Skorohod problem is Lipschitz continuous with applications", *Stochastics.* **35** 31-62.

[12] Einmahl, U. (1989). "Extensions of results of Komlós, Major and Tusnády to the Multivariate Case", *J. Multivariate Anal.* **28** 20-68.

[13] Flores, C. (1985). "Diffusion approximations for computer communications networks", in B. Gopinath (ed.), *Computer Communications, Proc. Symp. Appl. Math.*, American Math. Society, 83-124.

[14] Glynn, P. W. (1990). "Diffusion approximations", in D.P. Heyman and M.J. Sobel (eds.), *Handbooks in Operations Research and Management Science, II: Stochastic Models*, North-Holland, Amsterdam.

[15] Gordon ,W. J., and Newell, G. F. (1967). "Closed queueing systems with exponential servers", *Operations Research* **15** 254-265.

[16] Harrison, J. M. (1985). *Brownian Motion and Stochastic Flow Systems*, Wiley.

[17] Harrison, J. M. (1988). "Brownian models of queueing networks with heterogeneous customer populations", in W. Fleming and P.L. Lions (eds.), *Stochastic Differential Systems, Stochastic Control Theory and Applications*, IMA Volume **10**, Springer-Verlag, 147-186.

[18] Harrison, J. M., and Nguyen, V. (1990). "The QNET method for two-moment analysis of open queueing networks", *Queueing Systems* **6** 1-32.

[19] Harrison, J. M., and Nguyen, V. (1992). "Brownian models of multiclass queueing networks: current status and open problems", to appear in *Queueing Systems.*

[20] Harrison, J.M. and Reiman, M.I., "Reflected Brownian motion on an orthant", *The Annals of Prob.* **9**, 302–308, 1981.

[21] Harrison, J. M., and Wein, L.M. (1990). "Scheduling networks of queues: heavy traffic analysis of a two-station closed network", *Operations Research* **38** 1052-1064.

[22] Harrison, J. M., and Williams, R. (1987). "Brownian models of open queueing networks with homogeneous customer populations", *Stochastics* **22** 77-115.

[23] Harrison, J. M., Williams, R., and Chen, H. (1990). "Brownian models of closed queueing networks", *Stochastics and Stochastic Reports* **29** 37-74.

[24] Horváth, L. (1992). "Strong approximations of open queueing networks", *Math. O.R.* **17** 487-508.

[25] Iglehart D. L., and Whitt, W. (1970a). "Multiple channel queues in heavy traffic, I", *Adv. Appl. Prob.* **2** 150-177.

[26] Iglehart D. L., and Whitt, W. (1970b). "Multiple channel queues in heavy traffic, II", *Adv. Appl. Prob.* **2** 355-364.

[27] Jackson, J. R. (1963). "Jobshop-like queueing systems", *Management Science* **10** 131-142.

[28] Johnson, D. P. (1983). *Diffusion approximations for optimal filtering of jump processes and for queueing networks*, Ph.D. Dissertation, University of Wisconsin.

[29] Kaspi, H. and Mandelbaum, A. (1989). "On the ergodicity of a closed queueing network", submitted for publication.

[30] Kelly, F. P. (1979). *Reversibility and Stochastic Networks*, Wiley.

[31] Kelly, F. P., and C.N. Laws. (1991). "Dynamic routing in open queueing networks", preprint.

[32] Kingman, J.F. C. (1965). "The heavy traffic approximation in the theory of queues", in W.L. Smith et al. (eds.), *Proc. Symp. on Congestion Theory*, Univ. of North Carolina Press, 137-159.

[33] Kleinrock, L. (1976). *Queueing Systems Vol. II: Computer Applications*, Wiley.

[34] Krichagina, E., Lou, S., Sethi, S. and Taksar, M. (1991). "Production control in a failure-prone manufacturing system: diffusion approximation and asymptotic optimality", submitted to *Ann. Appl. Prob.*

[35] Lemoine, A. J. (1978). "Network of queues - a survey of weak convergence results", *Management Science* **24** 1175-1193.

[36] Meyn, S.P. and D. Down. (1993). "Stability of generalized Jackson networks", *Ann. Appl. Prob.*.

[37] Nguyen, V. (1993). "Fluid and diffusion approximations of a two-station mixed queueing network", Working paper WP# 3519-93-MSA, Sloan School, MIT.

[38] Reiman, M. I. (1984). "Open queueing networks in heavy traffic", *Math. of O.R.* **9** 441-458.

[39] Wein, L. M. (1990). "Scheduling networks of queues: heavy traffic analysis of a two-station network with controllable inputs", *Operations Research* **38** 1065-1078.

[40] Whitt, W. (1974). "Heavy traffic theorems for queues: a survey", in A.B. Clarke (ed.), *Mathematical Methods in Queueing Theory*, Springer Verlag.

[41] Whitt, W. (1980). "Some useful functions for functional limit theorems", *Math. of O.R.* **5** 67-85.

[42] Whitt, W. (1994). "An interesting example of a multiclass open queueing network", *Management Science*.

[43] Whittle, P. (1967). "Nonlinear migration processes", *Bull. Inst. Internat. Statist.* **42** 642-647.

[44] Whittle, P. (1968). "Equilibrium distributions for an open migration process", *J. Appl. Prob.* **5** 567-571.

[45] Yang, P. (1988). *Pathwise solutions for a class of linear stochastic systems*, Ph.D. dissertation, Department of Operations Research, Stanford University.

[46] Zhang, H., G. Hsu, and R. Wang. (1990). "Strong approximations for multiple channel queues in heavy traffic" *J. Appl. Prob.* **28**, 658-670.

[47] Zhang, H. (1990). "Strong approximations for open networks in heavy traffic", preprint.

[48] Zhang, H. (1993). "Strong approximations for irreducible closed queueing networks", preprint.