Heavy-Traffic Limits for Stationary Network Flows

Ward Whitt[1] and Wei You[2]

[1]Department of IEOR, Columbia University, New York, NY {ww2040@columbia.edu}
[2]Department of IEDA, HKUST, Hong Kong {weiyou@ust.hk}

October 19, 2019

Abstract

This paper studies the stationary customer flows in an open queueing network. The flows are the processes counting customers flowing from one queue to another or out of the network. We establish the existence of unique stationary flows in generalized Jackson networks and convergence to the stationary flows as time increases. We establish heavy-traffic limits for the stationary flows, allowing an arbitrary subset of the queues to be critically loaded. The heavy-traffic limit with a single bottleneck queue is especially tractable because it yields limit processes involving one-dimensional reflected Brownian motion. That limit plays an important role in our new nonparametric decomposition approximation of the steady-state performance using indices of dispersion and robust optimization.

1 Introduction

In this paper, we establish heavy-traffic limits for the stationary flows in a non-Markov open queueing network (OQN). By *flows*, we mean the departure processes, flows from one queue to another, superpositions of such processes and thus the internal arrival processes. We consider an OQN with K single-server stations, unlimited waiting space, and the first-come first-served service discipline. We assume that we have mutually independent renewal external arrival processes, sequences of independent and identically distributed (i.i.d.) service times and Markovian routing. Such a system is called a *generalized Jackson network* (GJN), because it generalizes the Markovian OQN analyzed by Jackson [17] in which all the interarrival times and service times have exponential distributions. Jackson OQN's are remarkably tractable because the vector of steady-state queue lengths (number in system) has a product-form distribution, just as if the queues were independent $M/M/1$ queues with the correct arrival rates.

The major theoretical advance for GJN's more general than Jackson OQN's has no doubt been the heavy-traffic limit theory [8, 19, 23] (which did not consider the flows). However, the practical application of that theory remains challenging, largely because the different queues in an OQN may have widely varying traffic intensities, with only a few being bottlenecks. The heavy-traffic limits can be extended to that case, as shown by Chen and Mandelbaum [6, 7], but there remains a need for effective numerical algorithms for computing performance measures, which properly account for a range of traffic intensities. See [11, 16] for previous algorithms.

Thus, early parametric-decomposition methods, as in [22], which treat the queues as mutually independent given a partial parametric specification of each internal arrival process (obtained by solving a system of linear equations) remain viable tools. This paper is part of our effort in [24, 25, 26, 28] to develop a new improved parametric-decomposition approximation for GJN's and more general OQN's, which have similar computational efficiency and ease of use.

Our main idea can perhaps best be seen by first considering a feed-forward GJN. Then the performance at each queue depends on the full model only through the service-time distribution at that queue and the arrival process to that queue. However, that arrival process tends to be relatively complicated, primarily because it tends to be non-renewal and depends on all the model parameters of the previous queues. In response, we partially characterize the stochastic properties of each stationary arrival process by its rate and index of dispersion for counts (IDC), which is a scaled version of the variance-time curve, a nonnegative real-valued function of time; see [25, 27].

To carry out this program, we need to do two things: (i) have a method to approximately solve for the steady-state performance of the general $G/GI/1$ queue, where the general stationary arrival process is partially characterized by the IDC and (ii) approximate the IDC of the arrival process at each queue. The first task is addressed in [25]. The present paper contributes to the second step.

For a feed-forward GJN, it is relatively easy to approximate the IDC of the arrival process at each queue, because the service times are independent of the arrival process. We can rely on the heavy-traffic limit for the stationary departure process in [24] and the functional central limit theorem (FCLT) of the superposition and splitting operations in §9.4 and §9.5 of [23].

The approximation of the IDC in a GJN with customer feedback is considerably more difficult, because of the correlation between the service times and the arrival processes. To develop such an approximation, we rely on the heavy-traffic limits for the flows established in this paper. For the full RQNA algorithm, including the extension to non-feed-forward OQN's, see [26, 28].

The heavy traffic limit for the flows in a GJN here extends the heavy-traffic limit for the stationary departure process in the $GI/GI/1$ model in [24]. That was evidently the first paper to establish a heavy-traffic limit for a stationary flow (other than an external arrival process) in a queueing model. Our main result in this paper is Theorem 3.1, which expresses a joint heavy-traffic limit for the centered flows with other processes. The limit for the flows is the final term in (3.13), which depends on the limits of other terms. However, Theorem 4.1 and Theorem 4.2 show that the limit simplifies dramatically when there is only a single bottleneck queue.

As before in [24], for our proof we rely heavily on the justification for interchanging the limits $t \to \infty$ and $\rho \uparrow 1$ in a GJN provided by Gamarnik and Zeevi [14] and Budharaja and Lee [5]. By allowing an arbitrary subset of the queues to be bottleneck queues (have nondegenerate limits), while the rest have null limits, we follow Chen and Mandelbaum [6, 7]. Even though the proofs follow quite directly from the existing literature, the asymptotic results here are evidently new.

As a preliminary step for our heavy-traffic limit, we establish conditions for the existence of stationary flows in a GJN and for convergence to those stationary flows as time evolves. For that we rely heavily on the Harris recurrence that was used to establish the stability of a GJN under appropriate regularity, as in Dai [9] (see the remark after Theorem 5.1 for earlier literature); also see Ch. VII of Asmussen [1].

The rest of the paper is organized as follows. We specify the model and establish the existence and convergence results (as time increases) for the stationary flows of a GJN in §2. We establish the main heavy-traffic limit for the stationary flows in §3. In §4 we treat the special case of a GJN with only one bottleneck queue, which is useful because it involves only one-dimensional RBM. We show that the approximation technique of feedback elimination discussed in §III of [22] is asymptotically correct in the HT limit. Finally, we draw conclusions in §5. Additional details for this paper appear in [28] and a longer version of this paper available on the authors' web pages.

2 The Stationary Flows in an Open Queueing Network

In this section, we establish the existence of unique stationary flows in a GJN and convergence to those stationary flows as time increases. These issues are complicated, but they are manageable under appropriate regularity conditions, in particular, if we construct a Markov process representation and make assumptions implying Harris recurrence as in §5 of [9], Chapter VII of [1], [14] and references there. In §2.1 we specify the model. Then in §2.2 we make assumptions implying the Harris recurrence and establish the existence, uniqueness and convergence result for the stationary flows.

2.1 The OQN Model

We start by formulating a general OQN model that goes beyond the assumptions we make to establish Harris recurrence. Let there be K single-server stations with unlimited waiting space and the first-come first-served (FCFS) discipline. We assume that the system starts empty at time 0, but that could be relaxed. We associate with each station i an external arrival point process $A_{0,i}$, which satisfies $A_{0,i}(t) < \infty$ with probability 1 for any t. Let $A_0 \equiv (A_{0,1}, \ldots, A_{0,K})$ denote the vector of all external arrival processes.

Let $\{V_i^l : l \geq 1\}$ denote the sequence of service times at station i and define the (uninterrupted) service point (counting) process as

$$S_i(t) = \max_{n \geq 0} \left\{ \sum_{l=1}^{n} V_i^l \leq t \right\}, \quad t \geq 0,$$

which we also assume to have finite sample path with probability 1.

In addition to external arrivals, departures from each station may be routed to other queues or out of the network. To specify the general routing (or splitting) process, let $\theta_i^l \in \{0,1\}^K$ indicate the routing vector of the l-th departure from queue i. Following standard conventions, at most one component of θ_i^l is 1, and $\theta_i^l = e_j$ indicates that the l-th departure from the i-th queue is routed to station j for $1 \leq j \leq K$, where e_j is the j-th standard basis of the Euclidean space $\mathbb{R}^K$. The case $\theta_i^l = 0$ indicates that the l-th departure from the i-th queue exits the system. The distbution of θ_i^l

is specified in Assumption 2.1. Finally, we define the routing decisions up to the n-th decision at station i by

$$\Theta_i(n) \equiv (\Theta_{i,1}(n), \ldots, \Theta_{i,K}(n)) \equiv \sum_{l=1}^{n} \theta_i^l,$$

and let $\Theta_{i,0}(n)$ denote the number of among the first n departing customers that exit the system from station i.

For the internal arrival flows, let $A_{i,j}$ be the customer flow from i to j. Each internal arrival flow $A_{i,j}$ splits from the departure process D_i according to the splitting decision process $\Theta_{i,j}$, so that

$$A_{i,j}(t) = \Theta_{i,j}(D_i(t)), \quad t \geq 0, \quad 1 \leq i \leq K, \quad 0 \leq j \leq K. \tag{2.1}$$

Let $A_{\text{int}}(t) \equiv (A_{i,j}(t) : 1 \leq i, j \leq K)$ denote the matrix of all internal arrival flows.

For total arrival process at station i, let

$$A_i(t) = A_{0,i}(t) + \sum_{j=1}^{K} A_{j,i}(t)$$

and let $A(t) \equiv (A_1(t), \ldots, A_K(t))$ be the vector of total arrival processes.

As observed in (7.1) and (7.2) in §7.2 of [6], the queue-length and departure processes at each queue are jointly uniquely characterized by the flow balance equations

$$Q_i(t) = Q_i(0) + A_i(t) - D_i(t)) \quad \text{and} \quad D_i(t) = S_i(B_i(t)), \quad t \geq 0, \quad 1 \leq i \leq K, \tag{2.2}$$

where $B_i(t)$ is the cumulative busy time of server i up to time t, which by work conservation satisfies

$$B_i(t) = \int_0^t 1_{Q_i(u)>0} du, \quad t \geq 0, \tag{2.3}$$

where 1_A is the indicator function with $1_A = 1$ on the set A and 0 elsewhere.

For the flow exiting the queueing system, let $D_{\text{ext},i}$ denote the flow that exits the system from station i. Hence

$$D_{\text{ext},i}(t) = \sum_{l=1}^{D_i(t)} \theta_{i,0}^l = \Theta_{i,0}(D_i(t)), \quad t \geq 0.$$

Finally, let $D_{\text{ext}}(t) \equiv (D_{\text{ext},1}(t), \ldots, D_{\text{ext},K}(t))$ be the vector of external departure processes.

2.2 Existence, Uniqueness and Convergence Via Harris Recurrence

In this section we establish the existence of unique stationary flows and convergence to them as time increases for any initial state. Toward that end, we make three assumptions, the first one being

Assumption 2.1 *We assume that the OQN is a GJN, in particular:*

(i) The K external arrival processes are mutually independent (possibly null) renewal processes with finite rates λ_i, where the interarrival times have finite squared coefficient of variation (scv, variance divided by the square of the mean) $c^2_{a_{0,i}}$ for $1 \leq i \leq K$.

(ii) The service times come from K mutually independent sequences of i.i.d. random variables with means $1/\mu_i$, $0 < \mu_i < \infty$, and finite scv $c^2_{s_i}$ for $1 \leq i \leq K$.

(iii) The routing is Markovian with a substochastic $K \times K$ routing matrix $P = (p_{i,j})_{1 \leq i,j \leq K}$ such that $p_{i,j} \geq 0$, $p_{i,0} \equiv 1 - \sum_{j=1}^K p_{i,j} \geq 0$ and $I - P'$ is invertible; For each $1 \leq i \leq K$, the sequence $\{\theta_i^1, \theta_i^2, \dots\}$ is i.i.d. with $P(\theta_i^l = e_j) = p_{i,j}$ and $P(\theta_i^l = 0) = p_{i,0} \equiv 1 - \sum_{j=1}^K p_{i,j}$.

(iv) The arrival, service and routing processes are mutually independent.

For completeness, we also assume that the network starts empty at time 0, so that no customer is in service or waiting, but this can be relaxed. The condition of finite scv's is used in the convergence of the distribution and in the next section; for relaxed assumptions, see the discussions below Theorem 2.1 and Theorem 2.2. Note that $I - P'$ is invertible if we assume that all customers eventually leave the system; see [8] or Theorem 3.2.1 of [18].

Let $U(t)$ denote the vector of residual external arrival times at time t; let $V(t)$ be the vector of residual service times at time t, set to 0 when the server is idle; and let the *system state process* be

$$\mathcal{S}(t) \equiv (Q(t), U(t), V(t)), \quad t \geq 0. \tag{2.4}$$

Under our assumption, the initial condition is specified by $\mathcal{S}(0) = (0, 0, 0)$. Since $\mathcal{S}$ is a piecewise-deterministic Markov process, the following result holds; see [9], which draws on [12].

Theorem 2.1 (strong Markov process) *Under Asusmption 2.1, the system state process $\mathcal{S}$ is a strong Markov process.*

To state the stability assumption, we let $\lambda_0 = (\lambda_{0,1}, \dots, \lambda_{0,K})$ be the external arrival rate vector and let $\lambda = (\lambda_1, \dots, \lambda_K)$ denote the vector of total arrival rate. We obtain λ by solving the *traffic-rate equations*

$$\lambda_i = \lambda_{0,i} + \sum_{i=1}^K \lambda_j p_{j,i}, \tag{2.5}$$

or, $(I - P')\lambda = \lambda_0$ in matrix form, where I denotes the $K \times K$ identity matrix and P' is the transpose of P. Let $\rho_i \equiv \lambda_i/\mu_i$ be the traffic intensity at station i.

Assumption 2.2 *The traffic intensities satisfy* $\max_i \rho_i < 1$.

Following convention, we say that the OQN is *stable* if the system state process in (2.4) is stable, i.e., if there exists a distribution π on $\mathbb{R}^{3K}$ for $\mathcal{S}(0)$ such that $\mathcal{S}(t)$ has that same distribution π for all $t \geq 0$. We now state the additional assumption to ensure the uniqueness of the stationary distribution π and the convergence of the distribution of $\mathcal{S}(t)$ to π.

Assumption 2.3 *Each non-null external arrival process has an interarrival-time distribution with a density that is positive for almost all* t.

Our assumption here implies the key assumption (A3) in both [9] and [10] that the distribution is unbounded and spread out, see also [9] and Chapter VII of [1]. This clearly avoids periodic behavior associated with the lattice case, but otherwise it is not restrictive for practical modeling.

The following theorem follows from Theorem 2 of [14] or Theorem 5.1 of [9] or Theorem 6.2 of [10], which extend earlier work on stability for OQNs in [3], [20] and [13].

Theorem 2.2 (existence, uniqueness and convergence) *Under Assumptions 2.1-2.3, the system state stochastic process* $\mathcal{S}$ *in* (2.4) *is a positive Harris recurrent Markov process. There exists a unique stationary distribution* π *and for every initial condition and the distribution of* $\mathcal{S}(t)$ *converges to* π *as* $t \to \infty$.

For a strong Markov process with right-continuous and left limit sample paths, the existence of a stationary distribution is shown in the early [2], which in turn draws on [15]. The uniqueness is shown in [9], which assumes that the interarrival times are unbounded, spreadout and have finite mean, and the service times have finite mean; see (1.2)-(1.5) there. The convergence in distribution follows from the convergence in total variation norm in Theorem 6.2 of [10], where they assumed finite $p+1$ moment for $p \geq 1$. Since our primary focus is the application to Robust Queue using the variance function, we content with the assumption of finite second moment, as in Assumption 2.1.

We now state the strong implications of Theorem 2.2, namely, the existence and convergence of stationary flows. Define the auxiliary cumulative process $\mathcal{C}$, as in §VI.3 of [1], by

$$\mathcal{C}(t) \equiv (B(t), Y(t)),$$

where $B_i(t)$ is the cumulative busy times for server i over interval $[0, t]$ and

$$Y_i(t) \equiv \mu_i(t - B_i(t)) \tag{2.6}$$

is the cumulative idle time of station i, scaled by the service rate μ_i.

To focus on the flows, we describe the GJN by the aggregate process

$$\mathcal{M}(t) \equiv (\mathcal{S}(t), \mathcal{C}(t), \mathcal{F}(t)), \tag{2.7}$$

where

$$\mathcal{F}(t) \equiv (A_0(t), A_{\mathrm{int}}(t), A(t), S(t), D(t), D_{\mathrm{ext}}(t)) \tag{2.8}$$

is a vector of cumulative point processes, with the processes defined in §2.1. We refer to $\mathcal{F}$ in (2.8) as the *flows*. We say that a flow is *stationary* if it has stationary increments. We refer to [21] and Chapter 6 of [4] for background on stationary stochastic processes and ergodicity.

Now, we consider the system that starts at time s. For the system state and auxillary processes, let $Q_s(t) = Q(s+t), U_s(t) = U(s+t)$, $V_s(t) = V(s+t)$, $B_s(t) = B(t+s) - B(s)$ and $Y_s(t) = Y(t+s) - Y(s)$, so that $\mathcal{S}_s \equiv (Q_s, U_s, V_s)$ is the system state process with initial condition $\mathcal{S}(s)$. For the flows, let $A_{0,s}(t) = A_0(t+s) - A_0(s)$ be the external arrival counting process that starts at time s. Similarly, let $A_{\mathrm{int},s}(t) = A_{\mathrm{int}}(t+s) - A_{\mathrm{int}}(s), A_s(t) = A(t+s) - A(s), D_s(t) = D(t+s) - D(s), D_{\mathrm{ext},s}(t) = D_{\mathrm{ext}}(t+s) - D_{\mathrm{ext}}(s)$ be the corresponding processes that starts at time s. The service processes $S_s(t)$ are more subtly defined by

$$S_{i,s}(t) \equiv S_i(B_i(s)+t) - S_i(B_i(s)), \quad \text{for} \quad i = 1, 2, \ldots, K, \tag{2.9}$$

which is a vector of delayed renewal processes with first intervals distributed as $V(s)$, the vector residual service time and at system time s (its i-th component is also the residual service time of the process S_i at time $B_i(s)$). Finally, let $\mathcal{C}_s \equiv (B_s, Y_s)$ and $\mathcal{F}_s \equiv (A_{0,s}, A_{\mathrm{int},s}, A_s, S_s, D_s, D_{\mathrm{ext},s})$.

Theorem 2.2 implies the existence and convergence of stationary flows.

Theorem 2.3 (Existence and convergence of the stationary flows) *Under Assumptions 2.1-2.3, there exists unique stationary and ergodic cumulative processes (with stationary increments satisfying the LLN)* $\mathcal{C}_e \equiv (B_e, Y_e), \mathcal{F}_e \equiv (A_{0,e}, A_{\mathrm{int},e}, A_e, S_e, D_e, D_{\mathrm{ext},e})$*, and a unique stationary process* $\mathcal{S}_e \equiv (Q_e, U_e, V_e)$*, such that, as* $s \to \infty$*,*

$$\mathcal{M}_s \equiv (\mathcal{S}_s, \mathcal{C}_s, \mathcal{F}_s) \Rightarrow (\mathcal{S}_e, \mathcal{C}_e, \mathcal{F}_e) \equiv \mathcal{M}_e, \tag{2.10}$$

where $\Rightarrow$ *denote weak convergence in each coordinate.*

3 Heavy-Traffic Limit Theorems for the Stationary Processes

To set the stage for our heavy-traffic limits, in §3.1 we present a centered representation of the flows. This representation parallels those used in [6, 7, 9, 19], but here we focus on the flows. Then in §3.2 we establish our main heavy-traffic limit.

3.1 Representation of the Centered Stationary Flows

Recall that the external arrival rate vector is λ_0, so the total arrival rates are given by $\lambda = (I - P')^{-1}\lambda_0$ as in (2.5). For service, we start with rate-1 base service process S_i^0 for station i and scale it by μ_i so that the service process at station i is denoted by $S_i \equiv S_i^0 \circ \mu_i e$ with $e(t) = t$ being the identity function. Let the center processes be defined by

$$\begin{aligned} &\tilde{A}_{0,i} = A_{0,i} - \lambda_{0,i} e, \tilde{A}_i = A_i - \lambda_i e, \tilde{D}_i = D_i - \lambda_i e, \\ &\tilde{\Theta}_{j,i} = \Theta_{j,i} \circ (S_j \circ B_j) - p_{j,i} S_j \circ B_j, \quad \text{and} \quad \tilde{S}_i = S_i \circ B_i - \mu_i B_i. \end{aligned} \tag{3.1}$$

Furthermore, let $X(t)$ be the *net-input process*, allowing the service to run continuously, defined as

$$X \equiv Q(t) - (I - P')Y, \tag{3.2}$$

where Y is defined in (2.6).

The next proposition expresses the queue length processes, the centered total arrival and the centered departure flows in terms of the centered external arrival, service and routing processes. Let ψ be the K-dimensional reflection map; e.g., see Chapter 14 of [23].

Proposition 3.1 (Centered representation) *The net-input process can be written as*

$$X = Q(0) + \tilde{A}_0 + \tilde{\Theta}'\mathbf{1} - (I - P')\tilde{S} + (\lambda_0 - (I - P')\mu)e, \tag{3.3}$$

while the queue length process can be written as

$$Q = X + (I - P')Y = \psi_{I-P'}(X), \tag{3.4}$$

where $\psi_{I-P'}$ is the K-dimensional reflection mapping with reflection matrix $I - P'$. In addition, the centered total arrival and departure processes can be written as

$$\tilde{A} = P'(I - P')^{-1}(Q(0) - Q) + (I - P')^{-1}\left(\tilde{A}_0 + \tilde{\Theta}'\mathbf{1}\right), \tag{3.5}$$

$$\tilde{D} = (I - P')^{-1}\left(Q(0) - Q + \tilde{A}_0 + \tilde{\Theta}'\mathbf{1}\right), \tag{3.6}$$

where the centered processes are defined in (3.1)*.*

Remark 3.1 (Stationary flows) *Note that the representation in Proposition 3.1 does not impose any assumption on the initial condition of the open queueing network. As ensured by Theorem 2.3, there exists a stationary distribution π such that the flows are stationary if $\mathcal{S}(0) \sim \pi$. With this specific initial condition, Proposition 3.1 applies to the stationary flows.*

3.2 Heavy-Traffic Limit with Any Subset of Bottlenecks

Throughout this section, we assume that the system is stationary in the sense of Theorem 2.3 and we suppress the subscript e to simplify the notation. We let an arbitrary pre-selected subset $\mathcal{H}$ of the K stations be pushed into the HT limit while other stations stay unsaturated. Two important special cases are: (i) $|\mathcal{H}| = K$ so that all stations approaches HT at the same time, which corresponds to the original case in [19]; and (ii) $|\mathcal{H}| = 1$ so that only one station is in HT. This second case is appealing for applications because the RBM is only one-dimensional. We focus on it in detail later.

To start, consider a family of systems indexed by ρ. Let the ρ-dependent service rates be

$$\mu_{i,\rho} \equiv \lambda_i/(c_i\rho), \quad 1 \le i \le K, \tag{3.7}$$

and set $c_i = 1$ for all $i \in \mathcal{H}$ and $c_i < 1$ for all $i \notin \mathcal{H}$. Equivalently, we have $\rho_i = c_i\rho$. For the pre-limit systems we have the same representation of the flows as described in Theorem 3.1, with the only exception that μ_i in (3.3) is now replaced by the ρ-dependent version in (3.7).

We now define the HT-scaled processes. As in the usual HT scaling, we scale time by $(1-\rho)^{-2}$ and scale space by $(1-\rho)$. Thus we make the definitions

$$\begin{aligned}
A^*_{0,i,\rho}(t) &\equiv (1-\rho)[A_{0,i}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_{0,i}t], \\
A^*_{i,\rho}(t) &\equiv (1-\rho)[A_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i t], \\
S^*_{i,\rho}(t) &\equiv (1-\rho)[S_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\mu_{i,\rho} t], \\
D^*_{i,\rho}(t) &\equiv (1-\rho)[D_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i t], \\
D^*_{\text{ext},i,\rho}(t) &\equiv (1-\rho)[D_{\text{ext},i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i p_{i,0} t], \\
A^*_{i,j,\rho}(t) &\equiv (1-\rho)[A_{i,j,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i p_{i,j} t], \\
\Theta^*_{i,j,\rho}(t) &\equiv (1-\rho)\left[\sum_{l=1}^{\lfloor (1-\rho)^{-2}t \rfloor} \theta^l_{i,j} - p_{i,j}(1-\rho)^{-2}t\right], \\
Q^*_{i,\rho}(t) &\equiv (1-\rho)Q_{i,\rho}((1-\rho)^{-2}t), \text{ for } 1 \le i,j \le K.
\end{aligned} \tag{3.8}$$

Furthermore, let $\Theta^*_{i,\rho} \equiv (\Theta^*_{i,j,\rho} : 1 \leq j \leq K)$; let $\Theta^*_{\text{ext},\rho} \equiv (\Theta^*_{i,0,\rho} : 1 \leq i \leq K)$; and let $\mathcal{F}^*_\rho$ collects all the scaled and centered flows, defined as

$$\mathcal{F}^*_\rho(t) \equiv (A^*_{0,\rho}(t), A^*_{\text{int},\rho}(t), A^*_\rho(t), S^*_\rho(t), D^*_\rho(t), D^*_{\text{ext},\rho}(t)). \tag{3.9}$$

Finally, let $Z^*_{i,\rho}(t) \equiv (1-\rho)Z_{i,\rho}((1-\rho)^2 t)$ denote the HT scaled workload process at station i in the ρ-th system.

Before presenting the HT limit of the systems, we introduce useful notation by discussing a modified system, that is asymptotically equivalent in heavy-traffic.

Remark 3.2 (Equivalent network) The system with bottleneck stations designated by $\mathcal{H}$ is asymptotically equivalent to a reduced $\mathcal{H}$-station network, where all non-bottleneck queues have zero service times. Equivalently, the non-bottleneck queues can be viewed as instantaneous switches. To obtain the rates and routing matrix in the equivalent network, we let $I_\mathcal{A}$ denote the $|\mathcal{A}| \times |\mathcal{A}|$ identity matrix for any index set $\mathcal{A}$; let $P_\mathcal{H}$ be the $|\mathcal{H}| \times |\mathcal{H}|$ submatrix of the original routing matrix P corresponding to the rows and columns in $\mathcal{H}$; let $P_{\mathcal{H}^c}$ be the submatrix of P corresponding to $\mathcal{H}^c$; and let $P_{\mathcal{H}^c,\mathcal{H}}$ collect the routing probabilities from stations in $\mathcal{H}^c$ to the ones in $\mathcal{H}$, similarly, define $P_{\mathcal{H},\mathcal{H}^c}$. Now the new routing matrix for the bottleneck stations, denoted by $\hat{P}_\mathcal{H}$, is

$$\hat{P}_\mathcal{H} = P_\mathcal{H} + P_{\mathcal{H},\mathcal{H}^c} \left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c,\mathcal{H}}. \tag{3.10}$$

Note that the inverse $\left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1}$ appearing in (3.10) is the fundamental matrix associated with the transient finite Markov chain with transition matrix $P_{\mathcal{H}^c}$. If we let $\hat{P}_{\mathcal{H}^c,\mathcal{H}}$ denote the matrix of the probabilities that the first visit to a bottleneck queue of an external arrival at a non-bottleneck queue $i \in \mathcal{H}^c$ is at $j \in \mathcal{H}$, then we have

$$\hat{P}_{\mathcal{H}^c,\mathcal{H}} = \sum_{l=0}^{\infty} (P_{\mathcal{H}^c})^l P_{\mathcal{H}^c,\mathcal{H}} = \left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c,\mathcal{H}}. \tag{3.11}$$

Similarly, for the new external arrival rate $\hat{\lambda}_{0,\mathcal{H}}$, we write

$$\hat{\lambda}_{0,\mathcal{H}} = \lambda_{0,\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} \lambda_{0,\mathcal{H}^c} = \lambda_{0,\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}} \left(I_{\mathcal{H}^c} - P'_{\mathcal{H}^c}\right)^{-1} \lambda_{0,\mathcal{H}^c}, \tag{3.12}$$

where $\lambda_{0,\mathcal{A}}$ denotes the column vector of the entries in λ_0 that corresponds to the index set $\mathcal{A}$. Since the total arrival rate in the modified system remains the same as the original system, we have

$$\hat{\lambda}_\mathcal{H} = (I - \hat{P}'_\mathcal{H})^{-1} \hat{\lambda}_{0,\mathcal{H}} = \lambda_\mathcal{H}.$$

To simplify notation, we suppress the subscript used in the identity matrix I in the rest of the paper whenever there is no confusion on its dimension. ∎

The following theorem states the joint heavy-traffic limit of the queue length process, the workload and waiting time processes, the splitting-decision process and all the flows. Combining conclusion (i) and (iii)-(v), we obtain explicit expression of the heavy-traffic limit of scaled and centered flows $\mathcal{F}^*_\rho$.

Theorem 3.1 (Heavy-traffic FCLT) *Under Assumption 2.1-2.3, consider a family of open queueing networks in stationarity, indexed by ρ. Let $\mathcal{H} \subset \{1, 2, \ldots, K\}$ denote the index of the bottleneck stations: Assume that $\mu_{i,\rho} = \lambda_i/(c_i\rho)$ for $1 \le i \le K$ and set $c_i = 1$ for all $i \in \mathcal{H}$ and $c_i < 1$ for all $i \notin \mathcal{H}$. Then, as $\rho \uparrow 1$,*

$$(Q^*_\rho, Z^*_\rho, \Theta^*_\rho, \Theta^*_{\text{ext},\rho}, \mathcal{F}^*_\rho) \Rightarrow (Q^*, Z^*, \Theta^*, \Theta^*_{\text{ext}}, \mathcal{F}^*), \tag{3.13}$$

where:

(i) For $0 \le i \le K$, $A^*_{0,i} = c_{a_{0,i}} B_{a_{0,i}} \circ \lambda_{0,i} e$ and $S^*_i = c_{s_i} B_{s_i} \circ \lambda_i e$, where $B_{a_{0,i}}$ and B_{s_i} are standard Brownian motions. $(\Theta^*, \Theta^*_{\text{ext}})$ is a zero-drift $(K+1)$-dimensional Brownian motion with covariance matrix $\Sigma_i = (\sigma^2_{jk} : 0 \le j, k \le K)$, where $\sigma^2_{j,j} = p_{i,j}(1-p_{i,j})\lambda_i$ and $\sigma^2_{j,k} = -p_{i,j}p_{i,k}\lambda_i$ for $0 \le i \ne j \le K$. Furthermore, $B_{a_{0,i}}$, B_{s_i} and $(\Theta^*, \Theta^*_{\text{ext}})$ are mutually independent, $1 \le i \le K$.

(ii) The limiting queue length process Q^* consists of two parts. $Q^*_{\mathcal{H}^c} \equiv 0$ and $Q^*_{\mathcal{H}}$ is a stationary $|\mathcal{H}|$-dimensional RBM

$$Q^*_{\mathcal{H}} \equiv \psi_{\mathcal{H}}\left(\hat{X}^*_{\mathcal{H}}\right),$$

where $\psi_{\mathcal{H}}$ is the $|\mathcal{H}|$-dimensional refelction map with reflection matrix $R_{\mathcal{H}} \equiv I - \hat{P}_{\mathcal{H}}$ and $\hat{X}^*_{\mathcal{H}}$ is a $|\mathcal{H}|$-dimensional Brownian motion

$$\hat{X}^*_{\mathcal{H}} = Q^*_{\mathcal{H}}(0) + \left(e'_{\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} e'_{\mathcal{H}^c}\right)\left(A^*_0 + (\Theta^*)'\,\mathbf{1}\right) - (I - \hat{P}_{\mathcal{H}})S^*_{\mathcal{H}} - \hat{\lambda}_{0,\mathcal{H}} e \tag{3.14}$$

where $e_{\mathcal{A}}$ collects columns in the K-dimensional identity matrix I that corresponds to index set $\mathcal{A}$; $\hat{P}_{\mathcal{H}}$, $\hat{P}_{\mathcal{H}^c,\mathcal{H}}$ and $\hat{\lambda}_{0,\mathcal{H}}$ are defined Remark 3.2; and $Q^*_{\mathcal{H}}(0)$ has unique stationary distribution of the stationary RBM.

(iii) The limiting total arrival process A^* is specified by

$$A^* = (I - P')^{-1}\left(A^*_0 + (\Theta^*)'\,\mathbf{1}\right) + P'(I - P')^{-1} e_{\mathcal{H}}\left(Q^*_{\mathcal{H}}(0) - Q^*_{\mathcal{H}}\right).$$

(iv) The limiting stationary departure process D^* is specified as

$$D^* = (I - P')^{-1} \left(Q^*(0) - Q^* + A_0^* + (\Theta^*)' \mathbf{1}\right).$$

In particular, $D^*_{\mathcal{H}^c} = Q^*_{\mathcal{H}^c} + A^*_{\mathcal{H}^c} - Q^*_{\mathcal{H}^c}(0) = A^*_{\mathcal{H}^c}$.

(v) The limiting internal arrival flow $A^*_{i,j}$ and external departure flow $D^*_{\mathrm{ext},i}$ can be expressed as

$$A^*_{i,j} = p_{i,j} D^*_i + \Theta^*_{i,j} \circ \lambda_i e, \quad \textit{and} \quad D^*_{\mathrm{ext},i} = p_{i,0} D^*_i + \Theta^*_{i,0} \circ \lambda_i e, \quad \textit{for} \quad 1 \le i, j \le K.$$

(vi) The limiting workload process is $Z^*_i = \lambda_i^{-1} Q^*_i$.

Proof. Much of the statement follows from [6, 7] and [5]. First, the HT limit for the state process with an arbitrary subset $\mathcal{H}$ of critically loaded stations follows from [6, 7]. Second, the HT limit for the steady-state queue length follows from [5]. The papers [14] and [5] do not consider non-bottleneck stations, but their arguments extend to that more general setting. (See Remark 3.3 below for discussion.) Because our basic model data involves only single arrival and service processes, with only the parameters being scaled, we do not need Assumption (A4) in [5]. We subsequently establish the heavy-traffic limits for the flows. We do so by exploiting the continuous mapping theorem with the direct representations of the stationary flows that we have established.

To carry out our proof, we work with the centered representation in Theorem 3.1, using the HT-scaling in (3.8). Thus, the HT-scaled net-input process is

$$X^*_\rho = Q^*_\rho(0) + A^*_{0,\rho} + \left(\tilde{\Theta}^*_\rho\right)' \mathbf{1} - (I - P')\tilde{S}^*_\rho + (\lambda_0 - (I - P')\mu_\rho)(1 - \rho)^{-1} e, \tag{3.15}$$

where $\tilde{S}^*_{i,\rho} \equiv S^*_{i,\rho} \circ \bar{\bar{B}}_{i,\rho}$, $\bar{\bar{B}}_{i,\rho} = (1-\rho)^2 B_{i,\rho} \circ (1-\rho)^{-2} e$, $\tilde{\Theta}^*_\rho$ is a matrix with its ij-th entry being $\Theta^*_{ij,\rho} \circ \overline{\overline{S \circ B}}_{i,\rho}$ and $\overline{\overline{S \circ B}}_\rho$ is a vector of length K with $\overline{\overline{S \circ B}}_{i,\rho} \equiv (1-\rho)^2 S_{i,\rho} \circ B_{i,\rho} \circ (1-\rho)^{-2} e$. The HT-scaled queue length can be written as $Q^*_\rho = X^*_\rho + (I - P')Y^*_\rho$. Re-writing $Q^*_{\mathcal{H},\rho}$ and $Q^*_{\mathcal{H}^c,\rho}$ in block-wise matrix representation yields

$$Q^*_{\mathcal{H},\rho} = \hat{X}^*_{\mathcal{H},\rho} + (I - \hat{P}'_{\mathcal{H}}) Y^*_{\mathcal{H},\rho}, \tag{3.16}$$

where $\hat{X}^*_{\mathcal{H},\rho} = X^*_{\mathcal{H},\rho} - P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q^*_{\mathcal{H}^c,\rho} - X^*_{\mathcal{H}^c,\rho})$.

Now, we substitute into $\hat{X}^*_{\mathcal{H},\rho}$ the expression for X^*_ρ from (3.15) in block matrix notation, leaving a constant $\hat{\eta}_\rho$ in the final deterministic drift term initially unspecified, to obtain

$$\hat{X}^*_{\mathcal{H},\rho} = Q^*_{\mathcal{H},\rho}(0) + \left(e'_{\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\right)\left(A^*_{0,\mathcal{H}^c,\rho} + (\tilde{\Theta}^*_\rho)'\mathbf{1}\right) + (I - \hat{P}'_{\mathcal{H}})\tilde{S}^*_{\mathcal{H},\rho}$$

$$+ P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q^*_{\mathcal{H}^c,\rho}(0) - Q^*_{\mathcal{H}^c,\rho}) + \hat{\eta}_\rho(1-\rho)^{-1}e. \tag{3.17}$$

To derive the drift term $\hat{\eta}_\rho = \lambda_0 - (I - P')\mu_\rho$, we re-write $\hat{\eta}_\rho$ into blocks

$$\hat{\eta}_\rho = \lambda_{0,\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\lambda_{0,\mathcal{H}^c} - (I - \hat{P}'_{\mathcal{H}})\mu_{\mathcal{H},\rho} = (I - \hat{P}'_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho}), \tag{3.18}$$

where the last equation follows from the block representation of the traffic-rate equation.

Now we are ready to deduce the claimed conclusions. First for conclusion (i), most follows directly from Donsker's theorem, Theorem 4.3.2 of [23], and the GJN assumptions. The exception is the limit

$$(\tilde{S}^*_\rho, \tilde{\Theta}^*_\rho) \Rightarrow (S^*, \Theta^*)$$

which follows from the continuous mapping theorem by a random-time-change argument as in [7].

For the convergence of the queue length process Q^*, we apply [5] to get

$$(Q^*_{\mathcal{H},\rho}(0), Q^*_{\mathcal{H}^c,\rho}(0)) \Rightarrow (Q^*_{\mathcal{H}}(0), Q^*_{\mathcal{H}^c}(0)) \quad \text{as} \quad \rho \uparrow 1.$$

In particular, we see that $Q^*_{\mathcal{H}^c,\rho} = 0$. For $Q^*_{\mathcal{H},\rho}$, we observe that $\hat{\eta}_\rho(1-\rho)^{-1}e \to -(I - \hat{P}_{\mathcal{H}})\lambda_{\mathcal{H}}e$ uniformly on bounded intervals and

$$Q^*_{\mathcal{H},\rho} = \hat{X}^*_{\mathcal{H},\rho} + (I - \hat{P}'_{\mathcal{H}})Y^*_{\mathcal{H},\rho} = \psi_{I-\hat{P}'_{\mathcal{H}}}(\hat{X}^*_{\mathcal{H},\rho}). \tag{3.19}$$

Conclusions (iii) and (iv) follow from the representations derived in Theorem 3.1, the continuous mapping theorem and the established convergence of the queue length process, the external arrival processes and the splitting-decision processes. To this end, we only need to apply diffusion scaling (accelerate time by $(1-\rho)^{-2}$ and scale space by $(1-\rho)$) to the representations in Proposition 3.1.

Next, conclusions (v) follows from the limit of the departure process and the FCLT of the splitting operation in §9.5 of [23]. Finally, the associated limits for the workload can be related to the limit for the queue length as indicated in [7] ■

Remark 3.3 (Elaboration on the application of [5]) We apply [5], but it must be extended to the model with non-bottleneck queues. We do not go through all details because we regard that step as minor, but we now briefly explain.

First, the main stability condition (A6) there holds in our setting here. The only difference is the use of ρ instead of n as in [5]. Comparing (3.8) here with (A5) there, for the bottleneck queues, the two scaling conventions are connected by setting $n = (1-\rho)^{-2}$, $\tilde{v}^n_i = 0$ and $\tilde{\beta}^n_i = -\lambda_i/\rho$. The stability condition here is then connected to that in [5] by setting $\theta_0 = -1$ in (13) there.

For the moment estimation in their Theorem 3.3, we treat $Q_{\mathcal{H}}$ and $Q^*_{\mathcal{H}^c}$ separately. For $Q_{\mathcal{H}}$, our representation (3.16) and (3.17) can be mapped to the representations (16) on p.51 of [5], but with slightly more complicated constant terms associated with the matrix multiplication we have in (3.17). Noting the expression of the drift term we have in (3.18), the rest of the proof is essentially the same. For $Q^*_{\mathcal{H}^c}$, by [6, 7], it is negligible in the sense of Theorem 3.3 of [5]. Theorem 3.4 of [5] relies only on the moment estimation as in their Theorem 3.3 and the strong Markov property of $\mathcal{S}(t)$ (which they denote as $X(t)$). Finally, Theorem 3.5 and Theorem 3.2 of [5] remain unchanged.

4 Examples

4.1 Functional Central Limit Theorem of the Flows

We now present important special cases of Theorem 3.1. We start with the case with no bottleneck queues. Suppose $|\mathcal{H}| = 0$ so that all stations are strictly non-bottleneck, i.e., $\mu_{i,\rho} = \lambda/(c_i\rho)$ where $c_i < 1$ for all i. As $\rho \uparrow 1$, the family of systems converges to a limiting system where the traffic intensity at station i is $\rho_i = c_i$. Hence, the scaling used in (3.8) corresponds to the diffusion scaling used in the usual FCLT. In particular, the diffusion limits can be written as

$$\begin{aligned} A^*_{0,i} &= c_{a_{0,i}} B_{a_{0,i}} \circ \lambda_{0,i} e, \quad S^*_i = c_{s_i} B_{s_i} \circ \lambda_i e, \\ A^* = D^* &= (I - P')^{-1} \left(A^*_0 + (\Theta^*)' \mathbf{1} \right), \\ A^*_{i,j} &= p_{i,j} D^*_i + \Theta^*_{i,j} \circ \lambda_i e, \quad D^*_{\text{ext},i} = p_{i,0} D^*_i + \Theta^*_{i,0} \circ \lambda_i e, \quad \text{for} \quad 1 \le i, j \le K. \quad \blacksquare \end{aligned}$$

4.2 Networks with One Bottleneck Queue

We now consider the special case in which there is only one bottleneck queue. Without loss of generality, let $\mathcal{H} = \{h\}$, so that station h is the only bottleneck station. This special case is especially tractable, because it involves one-dimensional RBM instead of multi-dimensional RBM. In particular, the limiting variance functions in such diffusion limits can be written explicitly. The variance functions are applied in RQNA [24, 25, 28]. We show that a feedback elimination procedure is asymptotically exact.

In doing so, we consider a reduced one-station network consist of the only bottleneck queue, while all non-bottleneck queues have service times set to 0 so that they serve as instantaneous switches. In the reduced network, we define an external arrival $\hat{A}_0$ to the bottleneck queue to be any external arrival that arrive at the bottleneck queue for the first time. Hence, an external arrival may have visited one or multiple non-bottleneck queues before its first visit to the bottleneck queue.

The following theorem implies that the reduced network is asymptotically equivalent to the original bottleneck queue in the sense of the stationary queue length process in the HT limit.

Theorem 4.1 *The HT limit $\hat{X}_h^*$ in (3.14), with $\mathcal{H} = \{h\}$, can be expressed as the following one-dimensional Brownian motion*

$$\hat{X}_h^* = Q_h^*(0) + \hat{A}^* + \left(\hat{\Theta}_S^* - (1-\hat{p})S_h^*\right) + \hat{\lambda}_{0,h}e, \tag{4.1}$$

where

$$\begin{aligned} \hat{A}^* &= A_{0,h}^* + \sum_{i\in\mathcal{H}^c}\left(\hat{p}_{i,h}A_{0,i}^* + \hat{\Theta}_{i,h}^*\right), \\ \hat{\Theta}_{i,h}^* &= \sqrt{\hat{p}_{i,h}(1-\hat{p}_{i,h})}B_{\hat{\Theta}_{i,h}} \circ \lambda_{0,i}e, \ \ \textit{and} \ \ \hat{\Theta}_S^* = \sqrt{\hat{p}(1-\hat{p})}B_{\hat{\Theta}_S} \circ \lambda_i e, \end{aligned} \tag{4.2}$$

while $\hat{p} \equiv \hat{P}_h$ and $\hat{p}_{i,h}$ is the (i,h)-th entry of $\hat{P}_{\mathcal{H}^c,\mathcal{H}}$ as in Remark 3.2; and $B_{\hat{\Theta}_{i,h}}, B_{\hat{\Theta}_S}$ are independent standard Brownian motions.

Furthermore, one can check by comparing Theorem 4.1 with Theorem 3.1 that (4.1) coincides with the HT limit of the net-input process in a single-server queue with feedback, where the external arrival process is $\hat{A}$, the service times remain unchanged and the feedback probability is $\hat{p}$.

As observed in Section III of [22], to develop effective parametric-decomposition approximations for OQNs it is often helpful to preprocess the model data by eliminating immediate feedback for queues with feedback. The immediate feedback returns the customer to the end of the line. The approximation step is to put the customer instead back at the head of the line, so as to receive all its (geometrically random number of) service times at once.

For our purpose here, we recognize all customers that feed back to the bottleneck queue as immediate feedback, even after visiting non-bottleneck queues. The probability of feedback is then exactly $\hat{p}$. After feedback elimination, the new service process $\hat{S}$ is the renewal process associated with the new service times, i.e., a geometric sum of the original service times at the bottleneck queue. Note that the modified service process after feedback elimination have a HT limit $\hat{S}^* \equiv \hat{\Theta}_S^* - (1-\hat{p})S_h^*$, where Θ_S^* is defined in (4.2). This matches exactly with the "service" component in (4.1). Hence, we have the following

Theorem 4.2 (Feedback elimination with one bottleneck queue) *For the bottleneck queue in the generalized Jackson network, consider the modified single-server queue with arrival process $\hat{A}$ and service process $\hat{S}$. The joint heavy-traffic limit for the queue length process, the waiting* *time process, the workload process and the external departure process in the original model can be expressed in terms of those in the modified system as*

$$(Q^*, Z^*, D^*_{\text{ext}}) \overset{dist.}{=} (\hat{Q}^*, (1-p)\hat{Z}^*, \hat{D}^*_{\text{ext}}).$$

5 Conclusions

After establishing existence and convergence (as time increases) for the stationary flows under Assumptions 2.1, 2.2 and 2.3 in Theorem 2.3, we established in Theorem 3.1 a general heavy-traffic limit for the system state process in (2.4) together with the flow process in (2.8), allowing an arbitrary subset of the stations to be critically loaded, while the rest are sub-critically loaded. For the heavy-traffic limit in Theorem 3.1, the processes of interest are centered and scaled as in (3.8) and (3.9). We then obtained explicit results for the special case in which zero or one station is critically loaded in §4.

There are many important topics for future research. First, it remains to establish an extension of Theorem 3.1 to the model generalized by allowing non-renewal external arrival processes, which requires generalizing the key supporting theorems in [5, 14]. It also remains to develop useful explicit formulas based on Theorem 3.1 when more than one station is critically loaded. Of course, it would also be good to obtain corresponding results for models with multiple classes and queues with multiple servers.

Acknowledgements

We thank Karl Sigman and Editor Sergey Foss for helpful discussion about Harris recurrence. This work was done while the Wei You was a graduate student at Columbia University, where both authors received support from NSF grant CMMI 1634133.

References

[1] S. Asmussen. *Applied Probability and Queues.* Springer, New York, second edition, 2003.

[2] J. Azema, M. Kaplan-Duflo, and D. Revuz. Invariant measures for classes of Markov processes (in french). *Probability Theory and Related Fields*, 8(3):157–181, 1967.

[3] A. A. Borovkov. Limit theorems for queueing networks, I. *Theory of Probability & Its Applications*, 31 (3):413–427, 1986.

[4] L. Breiman. *Probability.* SIAM, Philadelphia, 1992. Reprint of 1968 book in Classics in Applied Mathematics.

[5] A. Budhiraja and C. Lee. Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Mathematics of Operations Research*, 34(1):45–56, 2009.

[6] H. Chen and A. Mandelbaum. Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.*, 16(2):408–446, 1991.

[7] H. Chen and A. Mandelbaum. Stochastic discrete flow networks: diffusion approximations and bottlenecks. *The Annals of Probability*, 19(4):1463–1519, 1991.

[8] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization.* Springer, New York, 2001.

[9] J. Dai. On the positive Harris recurrence for multiclass queueing networks. *Ann Appl Probab*, 5:49–77, 1995.

[10] J. Dai and S. P. Meyn. Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control*, 40(11):1889–1904, 1995.

[11] J. Dai, V. Nguyen, and M. I. Reiman. Sequential bottleneck decomposition: an approximation method for generalized Jackson networks. *Operations research*, 42(1):119–136, 1994.

[12] M. H. A. Davis. Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic processes. *J. Roy. Stat.Soc. B*, 46(3):353–388, 1984.

[13] S. Foss. Ergodicity of queueing networks. *Siberian Math. J.*, 32:183–202, 1991.

[14] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Advances in Applied Probability*, 16(1):56–90, 2006.

[15] T. E. Harris. The existence of stationary measures for certain Markov processes. In *Proc. Third Berkeley Symp. Prob. and Stat.*, volume 2, pages 113–124. University of California, Berkely, CA, 1956.

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

[17] J. R. Jackson. Networks of waiting lines. *Operations Research*, 5(4):518–521, 1957.

[18] J. G. Kemeny and J. L. Snell. *Finite Markov Chains.* Springer, New York, 1976.

[19] M. I. Reiman. Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9(3):441–458, 1984.

[20] K. Sigman. The stability of open queueing networks. *Stochastic Processes and their Applications*, 35 (1):11–25, 1990.

[21] K. Sigman. *Stationary Marked Point Processes: An Intuitive Approach.* Chapman and Hall/CRC, New York, 1995.

[22] W. Whitt. The queueing network analyzer. *Bell Laboratories Technical Journal*, 62(9):2779–2815, 1983.

[23] W. Whitt. *Stochastic-Process Limits.* Springer, New York, 2002.

[24] W. Whitt and W. You. Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function. *Stochastic Systems*, 8(2):143–165, 2018.

[25] W. Whitt and W. You. Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research*, 66(1):184–199, 2018.

[26] W. Whitt and W. You. A robust queueing network analyzer based on indices of dispersion. working paper, Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2019.

[27] W. Whitt and W. You. The advantage of indices of dispersion in queueing approximations. *Operations Research Letters*, 7:99–104, 2019.

[28] Wei You. *A Robust Queueing Network Analyzer Based on Indices of Dispersion.* PhD thesis, Columbia University, 2019.