Heavy-Traffic Limits for Stationary Network Flows

Ward Whitt and Wei You

Department of Industrial Engineering and Operations Research,
Columbia University, New York, NY, 10027 {ww2040,wy2225@columbia.edu}

June 23, 2019

Abstract

This paper studies the stationary customer flows in 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 robust queueing.

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 [16] 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 heavy traffic limits here extend the heavy-traffic limit for the stationary departure process in the $GI/GI/1$ model in [23]. As before, 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 [6]. 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 [7, 8]. Even though the proofs follow quite directly from the existing literature, the asymptotic results here are evidently new. They play an important role in our new robust queueing network analyzer (RQNA) in [23, 24, 25], which is a nonparametric decomposition approximation.

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 [10] (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 [21] is asymptotically correct in the HT limit. Finally, we draw conclusions in §5.

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 [10], 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. 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 customers that exit the system from station i in the first n departures.

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 [7], 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 $c_{s_i}^2$ for $1 \le i \le K$.*

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

(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 [9] or Theorem 3.2.1 of [17].

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

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

Under our assumption, the initial condition is specified by $\mathcal{S}(0) = (0, 0, 0)$. The system state process $\mathcal{S}$ in (2.4) is an element of the function space $\mathcal{D}([0, \infty), \mathbb{R}^{3K})$ of real-valued functions on the half-line $[0, \infty)$ taking values in the Euclidean space $\mathbb{R}^{3K}$ that are right-continuous with left limits. As stated in §2.2 of [10], which draws on [12], Assumption 2.1 implies some basic regularity conditions.

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

We remark that Assumption 2.1 is stronger than needed to ensure the strong Markov property. Since $\mathcal{S}$ is a piecewise-deterministic Markov process (defined in §3 of [12]), §4 of [12] showed that if the expected number of jumps on any interval $[0, t]$ is finite, then the process possesses the strong Markov property.

We now state the stability assumption in the sense of the traffic intensities. 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*

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

or, in matrix form,

$$(I - P')\lambda = \lambda_0,$$

where I denotes the $K \times K$ identity matrix and P' is the transpose of P. Let $\lambda_{i,j} \equiv \lambda_i p_{i,j}$ be the rate of the internal arrival flow from i to j. Finally, 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 [10] and [11] that the distribution is unbounded and spread out, see also [10] 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 [10] or Theorem 6.2 of [11], which extend earlier work on stability for OQNs in [4], [19] 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 [10], 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 follows from [11] under the additional assumption of finite second moment.

We now state the strong implications of Theorem 2.2. For that, we consider the system that starts at time s. For the system state processes, let $Q_s(t) = Q(s+t), U_s(t) = U(s+t)$ and $V_s(t) = V(s+t)$, so that $\mathcal{S}_s \equiv (Q_s, U_s, V_s)$ is the system state process with initial condition $\mathcal{S}(s)$. Let $\Rightarrow$ denote weak convergence. Theorem 2.2 implies that

Corollary 2.1 *Under Assumptions 2.1-2.3, $Q_s(t)$ has unit (± 1) jumps and*

$$\mathcal{S}_s \Rightarrow \mathcal{S}_e \equiv (Q_e, U_e, V_e), \quad as \quad s \to \infty, \tag{2.6}$$

where $\mathcal{S}_e$ is the system state process with initial condition $\mathcal{S}_e(0)$ distributed as the stationary distribution π and $\Rightarrow$ denote weak convergence in each coordinate.

Proof. Assumption 2.3 implies that with probability 1, there is at most 1 (internal or external) arrival at any station and that the arrival times do not coincide with departure times at any station. Hence, Q_s only has unit-jumps.

From Theorem 2.2, we have the convergence of one-dimensional distribution

$$\mathcal{S}_s(t_1) \Rightarrow \mathcal{S}_e(t_1), \quad \text{for all} \quad t_1 \geq 0.$$

To extend the convergence to any finite-dimensional distribution, we utilize the Markov property of $\mathcal{S}(t)$ in Theorem 2.1. For any $t_2 = t_1 + \delta_1 > t_1$, the conditional probability distribution of the state $\mathcal{S}(t_1)$, conditioning on the past values up to the time t_1, depends only on the current state $\mathcal{S}_s(t_1)$. Apply Theorem 2.2 again with initial state $\mathcal{S}_s(t_1)$, we have

$$(\mathcal{S}_s(t_1), \mathcal{S}_s(t_2)) \Rightarrow (\mathcal{S}_e(t_1), \mathcal{S}_e(t_2)), \quad \text{for all} \quad 0 \leq t_1 < t_2.$$

By induction, the convergence can be extended to any finite-dimensional distribution. The weak convergence of the process $\mathcal{S}_s$ then follows from Theorem 12.6 in [3]. ∎

Now, we turn to the existence of stationary flows. Define the auxiliary cumulative process $\mathcal{C}$, as in §VI.3 of [1], by

$$\mathcal{C}(t) \equiv (B(t), Y(t)), \tag{2.7}$$

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

$$Y_i(t) \equiv \mu_i(t - B_i(t)) \tag{2.8}$$

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

$$\mathcal{M}(t) \equiv (\mathcal{S}(t), \mathcal{C}(t), \mathcal{F}(t)), \tag{2.9}$$

where

$$\mathcal{F}(t) \equiv (A_0(t), A_{\text{int}}(t), A(t), S(t), D(t), D_{\text{ext}}(t)) \tag{2.10}$$

is a vector of cumulative point processes, with the processes defined in §2.1. We refer to $\mathcal{F}$ in (2.10) as the *flows*. We say that a flow is *stationary* if it has stationary increments. We refer to [20] and Chapter 6 of [5] for background on stationary stochastic processes and ergodicity.

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 $A_{\text{int},s}(t) = A_{\text{int}}(t+s) - A_{\text{int}}(s), A_s(t) = A(t+s) - A(s), D_s(t) = D(t+s) - D(s), D_{\text{ext},s}(t) = D_{\text{ext}}(t+s) - D_{\text{ext}}(s)$, $B_s(t) = B(t+s) - B(s)$ and $Y_s(t) = Y(t+s) - Y(s)$ be the corresponding processes that starts at time s. The service processes $S_s(t)$ are more subtly defined by

$$S_{i,s}(t) \equiv S_i(B_i(s)+t) - S_i(B_i(s)), \quad \text{for} \quad i = 1, 2, \ldots, K, \tag{2.11}$$

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 $B_i(s)$). This definition of the service process allow us to write the departure process as a composition of the two processes S_s and B_s via

$$\begin{aligned} D_s(t) \equiv D(s+t) - D(s) &= (S \odot B)(s+t) - (S \odot B)(s) \\ &= (S_s \odot B_s)(t), \quad t \geq 0., \end{aligned} \tag{2.12}$$

where $\odot$ is understood as component-wise composition, i.e. $D_{i,s} = S_{i,s} \circ B_{i,s}$ for all i. Finally, let $\mathcal{C}_s \equiv (B_s, Y_s)$ and $\mathcal{F}_s \equiv (A_{0,s}, A_{\text{int},s}, A_s, S_s, D_s, D_{\text{ext},s})$.

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)*

$$\mathcal{C}_e \equiv (B_e, Y_e), \quad \mathcal{F}_e \equiv (A_{0,e}, A_{\text{int},e}, A_e, S_e, D_e, D_{\text{ext},e})$$

and a unique stationary process

$$\mathcal{S}_e \equiv (Q_e, U_e, V_e),$$

such that, as $s \to \infty$,

$$\mathcal{M}_s \equiv (\mathcal{S}_s, \mathcal{C}_s, \mathcal{F}_s) \Rightarrow (\mathcal{S}_e, \mathcal{C}_e, \mathcal{F}_e) \equiv \mathcal{M}_e, \tag{2.13}$$

where $\Rightarrow$ denote weak convergence in each coordinate. Furthermore, $A_{0,e}$ is the vector of equilibrium external arrival renewal processes, S_e is a vector of delayed renewal process with first interval distributed as $V_e(0)$.

Proof By Corollary 2.1 and the definition of S_s in (2.11), the convergence of $V_s(0) = V(s)$ implies the convergence of S_s to S_e, with the later one being a delayed renewal process with first interval distributed as $V_e(0)$ and other intervals distributed as a generic service time. Similarly, the components of $A_{0,s}$ are delayed renewal process with the first interval distributed as the components of $U_s(0)$, which is converging to the vector $A_{0,e}$ of the equilibrium external arrival processes. By the convergence of $\mathcal{S}_s$, we have as $s \to \infty$

$$(Q_s, U_s, V_s, A_{0,s}, S_s) \Rightarrow (Q_e, U_e, V_e, A_{0,e}, S_e) \tag{2.14}$$

We now turn our focus to the cumulative busy time process defined in (2.3). Let $h : \mathbb{R}_+ \to \mathbb{R}$ be a continuous function defined by $h(t) = t \wedge 1 \equiv \min\{t, 1\}, t \geq 0$. Then the busy period process can be written as

$$B_{i,s}(t) = \int_s^{s+t} 1_{Q_i(u)>0} du = \int_0^t 1_{Q_{i,s}(u)>0} du = \int_0^t h(Q_{i,s}(u)) du, \quad \text{for} \quad 1 \leq i \leq K. \tag{2.15}$$

The busy-period process thus has stationary increments because it is a measurable integrable function of $Q_{i,e}$, which is itself stationary. (Recall that general measurable functions of stationary process are stationary; see Proposition 6.6 of [5].) Let $\mathcal{C}(\mathbb{R}_+, \mathbb{R})$ denote the space of bounded continuous functions from $\mathbb{R}_+$ to $\mathbb{R}$, equipped with uniform norm. The mapping defined in (2.15) is a continuous mapping from $\mathcal{D}$ to $\mathcal{C}(\mathbb{R}_+, \mathbb{R})$; see Theorem 11.5.1 in [22]. The continuous mapping theorem then asserts that $B_s \Rightarrow B_e$, where $B_{i,e}(t) \equiv \int_0^t h(Q_{i,e}(u)) du$ for $t \geq 0$ and all i. For the cumulative idle-time process $Y_{i,s}(t) \equiv Y_i(t+s) - Y_i(s) = \mu_i(t - B_{i,s}(t))$, we note that t and $B_{i,s}(t)$ have continuous sample path, so that the linear function in (2.8) is continuous. Hence, we can extend the convergence as $s \to \infty$ in (2.14) to

$$(Q_s, U_s, V_s, A_{0,s}, S_s, B_s, Y_s) \Rightarrow (Q_e, U_e, V_e, A_{0,e}, S_e, B_e, Y_e) \tag{2.16}$$

The convergence established so far now implies associated convergence for the flows because the flow process $\mathcal{F}_s$ is determined by the state process $\mathcal{S}_s$. To make the connection, we introduce random vectors (T_s, J_s), where T_s is the time of the first jump in Q_s and J_s is the type of jump (external arrival to queue i, flow from queue i to queue j, or external departure from queue i), defined by

$$\begin{aligned} T_s &\equiv \min\{T_s^a, T_s^d\}, \quad \text{where} \\ T_s^a &\equiv \min\{U_{s,i}(0) : 1 \leq i \leq K\} \quad \text{and} \\ T_s^d &\equiv \min\{V_{s,i}(0) : Q_i(0) > 0, 1 \leq i \leq K\}. \end{aligned} \tag{2.17}$$

while $J_s = (0,i)$, (i,j) or $(i,0)$ if the minimum in the definition of T_s is attained, respectively, by T_s^a with index i, T_s^d with index i and the routing is to j, T_s^d with index i and the routing is to outside the network.

We observe that the we can regard $(T,J) : (s,\mathcal{S}_s) \to \mathbb{R} \times \mathcal{N}$, where $\mathcal{N}$ is a finite set, as a continuous map, so that $(T_s, J_s) \Rightarrow (T_e, J_e)$ as $s \to \infty$. We also observe that T_s is a stopping time with respect to the strong Markov process $\{\mathcal{S}_s(t) : t \geq 0\}$, so that we can repeat the construction for all successive jumps after time T_s.

In this way, we get convergence of the process of successive jump times and jump types (indexed by k)

$$\{(T_{s,k}, J_{s,k}) : k \geq 1\} \Rightarrow \{(T_{e,k}, J_{e,k}) : k \geq 1\} \quad \text{in} \quad (\mathbb{R} \times \mathcal{N})^{\infty} \quad \text{as} \quad s \to \infty. \tag{2.18}$$

That in turn implies convergence for the associated flow counting processes by applying the inverse map in §13.6 of [22] as stated. For example, we can write

$$N_s(t) \equiv \min\{k \geq 0 : T_{s,1} + \cdots + T_{s,k} \leq t\} \quad \text{and}$$
$$A_{s,i,j}(t) = \sum_{k=1}^{N_s(t)} 1_{J_{s,k}=(i,j)}. \quad \blacksquare$$

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 [7, 8, 10, 18], 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

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

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 (2.8).

The next theorem 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 [22].

Theorem 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 Theorem 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, Theorem 3.1 applies to the stationary flows.*

Proof With the standard flow conservation law, we can write the queue length process in terms of the centered processes

$$\begin{aligned} Q_i &= Q_i(0) + A_i - S_i \circ B_i \\ &= Q_i(0) + A_{0i} + \sum_{j=1}^{K} \Theta_{ji}(S_j \circ B_j) - S_i \circ B_i \\ &= Q_i(0) + (A_{0i} - \lambda_{0i} e) + \sum_{j=1}^{K} (\Theta_{ji}(S_j \circ B_j) - p_{ji} S_j \circ B_j) \\ &\quad - \sum_{j=1}^{K} (\delta_{ji} - p_{ji})(S_j \circ B_j - \mu_j B_j) + \sum_{j=1}^{K} (\delta_{ji} - p_{ji}) \mu_j (e - B_j) \end{aligned}$$

$$+\lambda_{0i}e-\sum_{j=1}^{K}(\delta_{ji}-p_{ji})\mu_j e.$$

Because $Y_i \equiv \mu_i\,(t-B_i)$ is the cumulative idle time, we can express Q in matrix form as

$$Q = Q(0)+A_0+\tilde{\Theta}'\mathbf{1}-(I-P')\tilde{S}+(I-P')Y+(\lambda_0-(I-P')\mu)e.$$

Furthermore, we have $Q = X+(I-P')Y$. Because Y is non-decreasing, $Y(0)=0$ and Y_i increases only when $Q_i=0$, (3.4) follows from the usual reflection argument.

Similarly, we can re-write the overall arrival process in terms of the centered processes

$$\begin{aligned}
A_i &= A_{0i}+\sum_{j=1}^{K}\Theta_{ji}(S_j\circ B_j)\\
&= (A_{0i}-\lambda_{0i}e)+\sum_{j=1}^{K}\left(\Theta_{ji}(S_j\circ B_j)-p_{ji}S_j\circ B_j\right)+\sum_{j=1}^{K}p_{ji}\left(S_j\circ B_j-\mu_jB_j\right)\\
&\quad -\sum_{j=1}^{K}p_{ji}\mu_j\left(e-B_j\right)+\lambda_{0i}e+\sum_{j=1}^{K}p_{ji}\mu_j e
\end{aligned}$$

or, in matrix notation, by

$$A=\tilde{A}_0+\tilde{\Theta}'\mathbf{1}+P'\tilde{S}-P'Y+(\lambda_0+P'\mu)e.$$

By (3.4), we have

$$\begin{aligned}
-P'Y &= P'(I-P')^{-1}(X-Q)\\
&= P'(I-P')^{-1}\left(Q(0)-Q+\tilde{A}_0+\tilde{\Theta}'\mathbf{1}+\lambda_0e\right)-P'\tilde{S}-P'\mu e.
\end{aligned}$$

Substituting into the matrix form of the arrival process, we have

$$\begin{aligned}
A &= \tilde{A}_0+\tilde{\Theta}'\mathbf{1}+P'\tilde{S}-P'Y+(\lambda_0+P'\mu)e\\
&= \tilde{A}_0+\tilde{\Theta}'\mathbf{1}+P'\tilde{S}+(\lambda_0+P'\mu)e\\
&\quad +P'(I-P')^{-1}\left(Q(0)-Q+\tilde{A}_0+\tilde{\Theta}'\mathbf{1}+\lambda_0e\right)-P'\tilde{S}-P'\mu e\\
&= P'(I-P')^{-1}\left(Q(0)-Q\right)+(I-P')^{-1}\left(\tilde{A}_0+\tilde{\Theta}'\mathbf{1}\right)+\lambda e. \qquad (3.7)
\end{aligned}$$

Finally, note that $D=Q(0)+A-Q$. ■

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 [18]; 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

$$\mu_{i,\rho} \equiv \lambda_i/(c_i\rho), \quad 1 \le i \le K, \tag{3.8}$$

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 (3.8).

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

$$\begin{aligned}
A^*_{0,i,\rho}(t) &\equiv (1-\rho)[A_{0,i}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_{0,i}t],\\
A^*_{i,\rho}(t) &\equiv (1-\rho)[A_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i t],\\
S^*_{i,\rho}(t) &\equiv (1-\rho)[S_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\mu_{i,\rho} t],\\
D^*_{i,\rho}(t) &\equiv (1-\rho)[D_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i t],\\
D^*_{\mathrm{ext},i,\rho}(t) &\equiv (1-\rho)[D_{\mathrm{ext},i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i p_{i,0} t],\\
A^*_{i,j,\rho}(t) &\equiv (1-\rho)[A_{i,j,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i p_{i,j} t],\\
\Theta^*_{i,j,\rho}(t) &\equiv (1-\rho)\left[\sum_{l=1}^{\lfloor (1-\rho)^{-2}t\rfloor} \theta^l_{i,j} - p_{i,j}(1-\rho)^{-2}t\right],\\
Q^*_{i,\rho}(t) &\equiv (1-\rho)Q_{i,\rho}((1-\rho)^{-2}t), \text{ for } 1 \le i,j \le K.
\end{aligned} \tag{3.9}$$

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

$$\mathcal{F}^*_\rho(t) \equiv (A^*_{0,\rho}(t), A^*_{\mathrm{int},\rho}(t), A^*_\rho(t), S^*_\rho(t), D^*_\rho(t), D^*_{\mathrm{ext},\rho}(t)). \tag{3.10}$$

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 and yet asymptotically equivalent system, where all service times at the nonbottleneck queues are set to zero.

Remark 3.2 (Equivalent network) This 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, so that they 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 $\mathcal{H}$; similarly, let $P_{\mathcal{H}^c}$ be the $|\mathcal{H}^c| \times |\mathcal{H}^c|$ submatrix of P corresponding to $\mathcal{H}^c$; and let $P_{\mathcal{H}^c,\mathcal{H}}$ collect the routing probablities from stations in $\mathcal{H}^c$ to the ones in $\mathcal{H}$, similarly, define $P_{\mathcal{H},\mathcal{H}^c}$. Now the new $|\mathcal{H}| \times |\mathcal{H}|$ routing matrix, denoted by $\hat{P}_{\mathcal{H}}$, is

$$\begin{aligned}\hat{P}_{\mathcal{H}} &= P_{\mathcal{H}} + \sum_{l=0}^{\infty} P_{\mathcal{H},\mathcal{H}^c} \left(P_{\mathcal{H}^c}\right)^l P_{\mathcal{H}^c,\mathcal{H}} \\ &= P_{\mathcal{H}} + P_{\mathcal{H},\mathcal{H}^c} \sum_{l=0}^{\infty} \left(P_{\mathcal{H}^c}\right)^l P_{\mathcal{H}^c,\mathcal{H}} \\ &= P_{\mathcal{H}} + P_{\mathcal{H},\mathcal{H}^c} \left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c,\mathcal{H}}. \qquad (3.11)\end{aligned}$$

Note that the inverse $\left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1}$ appearing in (3.11) 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

$$\hat{P}_{\mathcal{H}^c,\mathcal{H}} = \sum_{l=0}^{\infty} (P_{\mathcal{H}^c})^l P_{\mathcal{H}^c,\mathcal{H}} = \left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c,\mathcal{H}}. \qquad (3.12)$$

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

$$\hat{\lambda}_{0,\mathcal{H}} = \lambda_{0,\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} \lambda_{0,\mathcal{H}^c} = \lambda_{0,\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}} \left(I_{\mathcal{H}^c} - P'_{\mathcal{H}^c}\right)^{-1} \lambda_{0,\mathcal{H}^c}, \qquad (3.13)$$

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

$$\hat{\lambda}_{\mathcal{H}} = (I - \hat{P}'_{\mathcal{H}})^{-1} \hat{\lambda}_{0,\mathcal{H}} = \lambda_{\mathcal{H}}. \qquad (3.14)$$

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. As in [7, 8], we allow an arbitrary subset of nodes to be bottleneck queues (critically loaded) while the rest are sub-critically loaded. To treat the stationary processes, we apply [14] and [6], extended to include non-bottleneck queues. 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 [6].

Theorem 3.2 (Heavy-traffic FCLT) *Under Assumption 2.1-2.2, consider a family of open queueing networks in stationarity, indexed by ρ. Let $\mathcal{H} \subset \{1, 2, \dots, K\}$ 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 $\rho \uparrow 1$,*

$$(Q^*_\rho, Z^*_\rho, \Theta^*_\rho, \Theta^*_{\text{ext},\rho}, \mathcal{F}^*_\rho) \Rightarrow (Q^*, Z^*, \Theta^*, \Theta^*_{\text{ext}}, \mathcal{F}^*), \tag{3.15}$$

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 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 the net-input process associated with the bottleneck queues, defined below. Furthermore, $Q^*_{\mathcal{H}}(0)$ has unique stationary distribution of the stationary RBM. $\hat{X}^*_{\mathcal{H}}$ is a $|\mathcal{H}|$-dimensional Brownian motion

$$\hat{X}^*_{\mathcal{H}} = Q^*_{\mathcal{H}}(0) + \left(e'_{\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} e'_{\mathcal{H}^c} \right) \left(A^*_0 + (\Theta^*)' \mathbf{1} \right) - (I - \hat{P}_{\mathcal{H}}) S^*_{\mathcal{H}} - \hat{\lambda}_{0,\mathcal{H}} e \tag{3.16}$$

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 in (3.11), (3.12) and (3.13), respectively.

(iii) The total arrival process A^* can be regarded as a stationary process, having stationary increments, specified by

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

(iv) The 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 internal arrival flow $A^*_{i,j}$ can be expressed as

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

and the external departure flow can be expressed as

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

(vi) $Z^*_i = \lambda_i^{-1} Q^*_i$.

Proof of Theorem 3.2 Much of the statement follows from [7, 8] and [6]. First, the HT limit for the state process with an arbitrary subset $\mathcal{H}$ of critically loaded stations follows from [7, 8]. Second, the HT limit for the steady-state queue length follows from [6]. The papers [14] and [6] do not consider non-bottleneck stations, but their arguments extend to that more general setting. (See Remark 3.3 below for discussion.) 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 (3.9). Thus, the HT-scaled net-input process is

$$X^*_\rho = Q^*_\rho(0) + A^*_{0,\rho} + \left(\tilde{\Theta}^*_\rho \right)' \mathbf{1} - (I - P') \tilde{S}^*_\rho + (\lambda_0 - (I - P')\mu_\rho)(1 - \rho)^{-1} e, \tag{3.17}$$

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

$$Q^*_\rho = X^*_\rho + (I - P') Y^*_\rho.$$

We now re-write $Q^*_{\mathcal{H},\rho}$ and $Q^*_{\mathcal{H}^c,\rho}$ in block-wise matrix representation as follows

$$Q^*_{\mathcal{H},\rho} = X^*_{\mathcal{H},\rho} + (I - P'_{\mathcal{H},\mathcal{H}}) Y^*_{\mathcal{H},\rho} - P'_{\mathcal{H}^c,\mathcal{H}} Y^*_{\mathcal{H}^c,\rho} \tag{3.18}$$

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

Solving for $Y^*_{\mathcal{H}^c,\rho}$ in (3.19) and substituting into (3.18), we have

$$Q^*_{\mathcal{H},\rho} = \hat{X}^*_{\mathcal{H},\rho} + (I - \hat{P}'_{\mathcal{H}})Y^*_{\mathcal{H},\rho}, \tag{3.20}$$

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 (3.17), in block matrix notation, leaving a constant $\hat{\eta}_\rho$ in the final deterministic drift term initially unspecified, to obtain

$$\begin{aligned}
\hat{X}^*_{\mathcal{H},\rho} = {} & Q^*_{\mathcal{H},\rho}(0) + A^*_{0,\mathcal{H},\rho} + e'_{\mathcal{H}}(\tilde{\Theta}^*_\rho)'\mathbf{1} - (I - P'_{\mathcal{H},\mathcal{H}})\tilde{S}^*_{\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}\tilde{S}^*_{\mathcal{H}^c,\rho} \\
& - P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}Q^*_{\mathcal{H}^c,\rho} \\
& + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\Big(Q^*_{\mathcal{H}^c,\rho}(0) + A^*_{0,\mathcal{H}^c,\rho} \\
& + e'_{\mathcal{H}^c}(\tilde{\Theta}^*_\rho)'\mathbf{1} - (I - P'_{\mathcal{H}^c,\mathcal{H}^c})\tilde{S}^*_{\mathcal{H}^c,\rho} + P'_{\mathcal{H},\mathcal{H}^c}\tilde{S}^*_{\mathcal{H},\rho}\Big) + \hat{\eta}_\rho(1-\rho)^{-1}e \\
= {} & Q^*_{\mathcal{H},\rho}(0) + A^*_{0,\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}A^*_{0,\mathcal{H}^c,\rho} + (I - \hat{P}'_{\mathcal{H}})\tilde{S}^*_{\mathcal{H},\rho} \\
& + e'_{\mathcal{H}}(\tilde{\Theta}^*_\rho)'\mathbf{1} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}e'_{\mathcal{H}^c}(\tilde{\Theta}^*_\rho)'\mathbf{1} \\
& + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q^*_{\mathcal{H}^c,\rho}(0) - Q^*_{\mathcal{H}^c,\rho}) + \hat{\eta}_\rho(1-\rho)^{-1}e.
\end{aligned}$$

Now we derive the drift term $\hat{\eta}_\rho$. To start, let

$$\eta_\rho = \lambda_0 - (I - P')\mu_\rho.$$

Just like how we treat the HT-scaled queue length process, we can re-write η_ρ into blocks

$$\eta_{\mathcal{H},\rho} = \lambda_{0,\mathcal{H}} - (I - P'_{\mathcal{H},\mathcal{H}})\mu_{\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}\mu_{\mathcal{H}^c,\rho}, \tag{3.21}$$

$$\eta_{\mathcal{H}^c,\rho} = \lambda_{0,\mathcal{H}^c} - (I - P'_{\mathcal{H}^c,\mathcal{H}^c})\mu_{\mathcal{H}^c,\rho} + P'_{\mathcal{H},\mathcal{H}^c}\mu_{\mathcal{H},\rho}. \tag{3.22}$$

Hence

$$\begin{aligned}
\hat{\eta}_\rho &\equiv \eta_{\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\eta_{\mathcal{H}^c,\rho} \\
&= \lambda_{0,\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\lambda_{0,\mathcal{H}^c} - (I - \hat{P}'_{\mathcal{H}})\mu_{\mathcal{H},\rho}. \qquad (3.23)
\end{aligned}$$

Note that the traffic-rate equation can be written as

$$\lambda_{0,\mathcal{H}} = (I - P'_{\mathcal{H},\mathcal{H}})\lambda_{\mathcal{H}} - P'_{\mathcal{H}^c,\mathcal{H}}\lambda_{\mathcal{H}^c},$$

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

Substitute both $\lambda_{0,\mathcal{H}}$ and $\lambda_{0,\mathcal{H}^c}$ into (3.23), we have

$$\hat{\eta}_\rho = (I - \hat{P}'_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho}). \tag{3.24}$$

To summarize, the HT-scaled net-input process associated with the bottleneck queues can be expressed as

$$\begin{aligned}\hat{X}^*_{\mathcal{H},\rho} \quad = \quad & Q^*_{\mathcal{H},\rho}(0) + A^*_{0,\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}A^*_{0,\mathcal{H}^c,\rho} - (I - \hat{P}'_{\mathcal{H}})\tilde{S}^*_{\mathcal{H},\rho} \\ & + e'_{\mathcal{H}}(\tilde{\Theta}^*_\rho)'\mathbf{1} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}e'_{\mathcal{H}^c}(\tilde{\Theta}^*_\rho)'\mathbf{1} \\ & + (I - \hat{P}_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho})(1-\rho)^{-1}e \\ & + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q^*_{\mathcal{H}^c,\rho}(0) - Q^*_{\mathcal{H}^c,\rho}).\end{aligned} \tag{3.25}$$

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 [22], 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 shown in [8].

For conclusion (ii), we apply [6] to get

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

Then the conclusion (ii) follows from Theorem 6.1 of [8]. In particular, there we see that $Q^*_{\mathcal{H}^c}$ is null, so that we can treat the two components of $(Q^*_{\mathcal{H},\rho}, Q^*_{\mathcal{H}^c,\rho})$ separately. First, to treat $Q^*_{\mathcal{H},\rho}$, we apply the continuous mapping theorem with the reflection map using the representation above. To do so, we observe that, as $\rho \uparrow 1$,

$$(I - \hat{P}_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho})(1-\rho)^{-1}e \to -(I - \hat{P}_{\mathcal{H}})\lambda_{\mathcal{H}}e$$

and

$$Q^*_{\mathcal{H},\rho} = \hat{X}^*_{\mathcal{H},\rho} + (I - \hat{P}'_{\mathcal{H}})Y^*_{\mathcal{H},\rho} = \psi_{I-\hat{P}'_{\mathcal{H}}}(\hat{X}^*_{\mathcal{H},\rho}). \tag{3.26}$$

Conclusions (iii) and (iv) follows 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 Theorem 3.1 so that

$$A^*_\rho = P'(I-P')^{-1}\left(Q^*_\rho(0) - Q^*_\rho\right) + (I-P')^{-1}\left(A^*_{0,\rho} + (\tilde{\Theta}^*_\rho)'\mathbf{1}\right),$$
$$D^*_\rho = (I-P')^{-1}\left(Q^*_\rho(0) - Q^*_\rho + A^*_{0,\rho} + (\tilde{\Theta}^*_\rho)'\mathbf{1}\right). \quad (3.27)$$

The second expression follows from the fact that $Q^*_{\mathcal{H}^c} = 0$.

Next, conclusions (v) follows from the limit of the departure process and the FCLT of the splitting operation in §9.5 of [22]. Finally, the associated limits for the workload can be related to the limit for the queue length as indicated in [8]. ∎

Remark 3.3 (Elaboration on the application of [6]) We apply [6], 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. Notice that our scaling convention here relies on the traffic intensity parameter ρ instead of the scaling parameter n used in [6]. Comparing (3.9) 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 [6] 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 (3.20) and (3.25) can be mapped to the representations (16) on p.51 of [6], but with slightly more complicated constant terms associated with the matrix multiplication we have in (3.25). Noting the expression of the drift term we have in (3.24), the rest of the proof is essentially the same. For $Q^*_{\mathcal{H}^c}$, by [7, 8], it is negligible in the sense of Theorem 3.3 of [6]. Theorem 3.4 of [6] relies only on the moment estimation as in their Theorem 3.3 and the strong Markov property of $\mathcal{S}(t)$ (which they denoted as $X(t)$). Finally, Theorem 3.5 and Theorem 3.2 of [6] remain unchanged.

Remark 3.4 (Functional Central Limit Theorem of the flows) An important special case of Theorem 3.2 arises when we set $|\mathcal{H}| = 0$ 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 (3.9) corresponds to the diffusion scaling used in the usual FCLT. In particular, the diffusion limits can be written as

$$A^*_{0,i} = c_{a_{0,i}} B_{a_{0,i}} \circ \lambda_{0,i} e,$$

$$
\begin{aligned}
S_i^* &= c_{s_i} B_{s_i} \circ \lambda_i e, \\
A^* = D^* &= (I - P')^{-1} \left(A_0^* + (\Theta^*)' \mathbf{1} \right), \\
A_{i,j}^* &= p_{i,j} D_i^* + \Theta_{i,j}^* \circ \lambda_i e, \\
D_{\text{ext},i}^* &= p_{i,0} D_i^* + \Theta_{i,0}^* \circ \lambda_i e, \quad \text{for} \quad 1 \le i, j \le K.
\end{aligned}
$$

where $B_{a_{0,i}}$ and B_{s_i} and $(\Theta_{i,j}^* : 0 \le j \le K)$ are Brownian motions defined as in part (i) of Theorem 3.2. ■

4 The Special Case of Only One Bottleneck Queue

In this section we consider the special case in which there is only one bottleneck queue, which 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 [23, 24, 25].

We start with the easiest special case: when $|\mathcal{H}| = K = 1$, which corresponds to the $GI/GI/1$ queue with i.i.d. customer feedback. We observe that this model is asymptotically equivalent to a modified single-server queue model without customer feedback, where the arrival process is generalized to include the immediate feedback.

Furthermore, we show that it is asymptotically correct in HT for a GJN with a single bottleneck queue to eliminate all feedback prior to analysis. We show how to quantify feedback elimination.

4.1 Single-Server Queue with Customer Feedback

Consider a single-server queue with customer feedback as depicted in Figure 1. Let A_0 denote the renewal external arrival process with rate λ_0 and scv $c_{a_0}^2$. Let the feedback probability be p, so that the effective arrival rate is $\lambda = \lambda_0/(1-p)$. Let service times be i.i.d. with rate $\mu_\rho = \lambda/\rho$ and scv c_s^2, hence a traffic intensity of ρ. Let A denote the total arrival process; let A_{int} be the feedback flow; let S denote the service process; let D be the total departure process; and let D_{ext} denote the flow that exits the system.

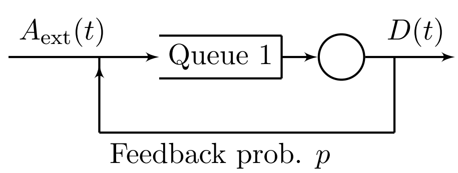

Figure 1: A single-server queue with feedback example.

As observed in Section III of [21], 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. Clearly this does not alter the queue length process and the workload process. The modified system does not have a feedback flow and the new service time will be the geometric random sum of the i.i.d. copies of the original service times, let $\tilde{S}$ denote the new service counting process.

This modification results in a change in the service rate and service scv. The new service rate is $(1-p)\mu = (1-p)\lambda/\rho = \lambda_0/\rho$ and, by conditional variance formula, the new scv is $\tilde{c}_s^2 = p+(1-p)c_s^2$. Hence, the heavy-traffic limit of the new service process is $\tilde{S}^* \equiv \tilde{c}_s^2 \tilde{B}_s \circ \lambda_0 e$. We now claim that $\tilde{S}^* \overset{dist.}{=} \Theta^* - (1-p)S^*$. To this end, note that $\Theta^* = \sqrt{p(1-p)} B_\Theta \circ \lambda e$ and $S^* = c_s B_s \circ \lambda e$, where B_Θ, B_s are independent standard Brownian motions (zero drift and unit variance) and $\lambda_0 = (1-p)\lambda$.

The joint HT limit for the flows in the original system can be obtained from Theorem 3.2 by setting $K = 1$ and $\mathcal{H}$ to be the only queue in the system (also the bottleneck queue). From part (ii) of Theorem 3.2, we have

$$X^* \overset{dist.}{=} Q^*(0) + A_0^* + \tilde{S}^* - \lambda_0 e. \tag{4.1}$$

Let $\tilde{Q}^*, \tilde{Z}^*$ denote the HT limit of the queue length process and the workload process in the modified single-server queue without feedback, having arrival process A_0 and service process $\tilde{S}$. Standard heavy-traffic theory implies that (4.1) is exactly the HT limit of the net-input process of a single-server queue so that $\tilde{Q}^* \overset{dist.}{=} Q^*$. Hence, we have

$$\tilde{Z}^* \equiv \lambda_0^{-1} \tilde{Q}^* \overset{dist.}{=} (1-p)^{-1} \lambda^{-1} Q^* \equiv (1-p)^{-1} Z^*.$$

Note that the expected number of visit for the same customer is $(1-p)^{-1}$. This implies that for approximating the waiting time and workload in the original system, we need to adjust for per-visit version by multiplying the values in the modified system by $(1-p)$.

Theorem 4.1 (Eliminating immediate feedback) *For the single-server queue with feedback model in Figure 1, consider the modified single-server queue, where immediate feedback are eliminated by placing the feedback customers at the head of the line. 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*

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

4.2 Networks with One Bottleneck Queue

We now consider the more general special case in which $K \geq 1$ but $|\mathcal{H}| = 1$ and show that feedback elimination is also asymptotically correct for networks with one bottleneck.

In doing so, we first observe that a GJN with one bottleneck queue that the bottleneck queue is asymptotically equivalent to a $G/GI/1$ single-server queue with feedback in the HT limit, where the arrival process is a complex superposition of renewal arrival processes. We derive the explicit expression for the external arrival process and feedback probability in the equivalent network.

We start with a convenient representation of the HT limit of the bottleneck queue. Without loss of generality, let $\mathcal{H} = \{h\}$, so that station h is the only bottleneck station. Let $\hat{p}_{i,h}$ be the (i,h)-th component of $\hat{P}_{\mathcal{H}^c,\mathcal{H}}$ in *(3.12)* and recall that $\hat{p} \equiv \hat{P}_h$ is the feedback probability defined in Remark 3.2.

Theorem 4.2 *The HT limit $\hat{X}^*_h$ in* (3.16)*, with $\mathcal{H} = \{h\}$, can be expressed as the following one-dimensional Brownian motion*

$$\hat{X}^*_h = Q^*_h(0) + \hat{A}^* + \left(\hat{\Theta}^*_S - (1-\hat{p})S^*_h\right) + \hat{\lambda}_{0,h}e, \tag{4.2}$$

where

$$\hat{A}^* = A^*_{0,h} + \sum_{i\in\mathcal{H}^c}\left(\hat{p}_{i,h}A^*_{0,i} + \hat{\Theta}^*_{i,h}\right), \tag{4.3}$$

$$\hat{\Theta}^*_{i,h} = \sqrt{\hat{p}_{i,h}(1-\hat{p}_{i,h})}B_{\hat{\Theta}_{i,h}} \circ \lambda_{0,i}e, \ \ \textit{and} \ \ \hat{\Theta}^*_S = \sqrt{\hat{p}(1-\hat{p})}B_{\hat{\Theta}_S} \circ \lambda_i e, \tag{4.4}$$

while $B_{\hat{\Theta}_{i,h}}$ and $B_{\hat{\Theta}_S}$ are independent standard Brownian motions.

Proof Since the drift term, the terms associated with A^*_0 and S^*_h remain unchanged, it suffices to show that the terms related with the splitting decision processes share the same variance. In fact, by algebraic manipulation, one can check that

$$\begin{aligned}\mathrm{Var}\left(\sum_{i\in\mathcal{H}^c}\hat{\Theta}^*_{i,h} + \hat{\Theta}^*_S\right) &= \sum_{i\in\mathcal{H}^c}\hat{p}_{i,h}(1-\hat{p}_{i,h})\lambda_{0,i}e + \hat{p}(1-\hat{p})\lambda_i e \\ &= \sum_{i=1}^{K}\left(e'_h + \hat{P}'_{\mathcal{H}^c,h}e'_{\mathcal{H}^c}\right)\Sigma_i\left(e_h + e_{\mathcal{H}^c}\hat{P}_{\mathcal{H}^c,h}\right)e \\ &= \mathrm{Var}\left(e'_h\left(\Theta^*\right)'\mathbf{1} + \hat{P}'_{\mathcal{H}^c,h}e'_{\mathcal{H}^c}\left(\Theta^*\right)'\mathbf{1}\right)\end{aligned}$$

where Σ_i are the variance matrix defined in Theorem 3.2. ■

Now, 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. In particular, the external arrival process can be expressed as the superposition of (i) the original external arrival process $A_{0,h}$ at station h; and (ii) the Markov splitting of the external arrival process $A_{0,i}$ at station i with probability $\hat{p}_{i,h}$, for $i \in \mathcal{H}^c$.

Theorem 4.2 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. Furthermore, one can check by comparing Theorem 4.2 with part (ii) of Theorem 3.2 that (4.2) 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 $\hat{p}$.

We then eliminate immediate feedback customers just as in Theorem 4.1, but with the extended interpretation of immediate feedback. Recalling that the non-bottleneck queues act as instantaneous switches, 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 $\hat{p} \equiv \hat{P}_h$ as in Remark 3.2. 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 (4.4), just as discussed in Section 4.1. This matches exactly with the "service" component in (4.2). Hence, we have the following extension of Theorem 4.1.

Theorem 4.3 (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*

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

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 3.2 a general heavy-traffic limit for the system state process in (2.4) together with the flow process in (2.10), 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 3.2, the processes of interest are centered and scaled as in (3.9) and (3.10). We then obtained explicit results for the special case in which only one station is critically loaded in §4.

There are many important topics for future research. First, it remains to establish an extension of Theorem 3.2 to the model generalized by allowing non-renewal arrival processes, which requires generalizing the key supporting theorems in [6, 14]. It also remains to develop useful explicit formulas based on Theorem 3.2 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. We 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] P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1999.

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

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

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

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

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

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

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

[11] 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.

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

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

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

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

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

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

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

[23] 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.

[24] 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.

[25] 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, 2018.