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}

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.

History: January 26, 2019; updated on December 20, 2019.

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.

1.1 The Flows in an OQN

The flows are special stochastic point processes, for which there is a well-developed general theory, as in [18, 19]. There also is a substantial literature on the general structure of stationary point processes in queueing systems, as in Chapter 1 of [3] and [42], but concrete results, such as explicit formulas describing the stochastic variability of the flows over time, are extremely rare. The familiar exception is the Markovian Jackson OQN analyzed by Jackson [33], for which there is a substantial theory, as in Ch. 4 of [45], but even in Markovian Jackson networks, the flows can be quite complicated. First, by reversibility, for Jackson networks, the departure processes out of the network from the queues are independent Poisson processes, but the internal flows need not be Poisson, even though the product-form property holds. In particular, the flows are Poisson if and

only if they are not part of a loop; see [36, 44]. For non-Markov open networks, the flows are even more complicated. As discussed in [17, 21] and references there, the stationary departure process from a $GI/GI/1$ queue is a renewal process (ordinary or stationary) if and only if the queue is an $M/M/1$ queue, in which case it is a Poisson process.

In this paper, 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 Jackson network in which all the interarrival times and service times have exponential distributions.

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 [49]. 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 3.2, which expresses a joint heavy-traffic limit for the centered flows with other processes. The limit for the flows is the final term in (3.11), which depends on the limits of other terms. However, Theorem 4.2 and Theorem 4.3 show that the limit simplifies dramatically when there is only a single bottleneck queue.

As before in [49], 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 [23] and Budharaja and Lee [8]. 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 [10, 11]. 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 [14] (see the remark after Theorem 5.1 for earlier literature); also see Ch. VII of Asmussen [1].

1.2 The Index of Dispersion (IDC) of a Stationary Point Pricess

In addition to contributing to a better understanding of the stationary flows in GJNs, we are particularly interested in the application of heavy-traffic limits in the approximation of key performance measures in non-Markov OQNs.

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. However, relatively little is known about the exact steady-state performance of a GJN. The major theoretical advance for GJN's more general than Jackson OQN's has no doubt been the heavy-traffic limit theory [12, 38, 48] (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 [10, 11], but there remains a need for effective numerical algorithms for computing performance measures, which properly account for a range of traffic intensities. See [16, 28] for previous algorithms.

This paper is part of our effort to develop a new improved non-parametric decomposition approach based on the indices of dispersion for counts (IDC) of the stationary flows [49, 50, 51, 53], which have similar computational efficiency and ease of use as QNA in [46].

As in §4.5 of [13], the IDC is a scaled version of the variance-time function; i.e., given a *stationary* arrival counting process $A(t)$ with rate λ, the IDC is the function

$$I_a(t) \equiv \frac{\mathrm{Var}(A(t))}{E[A(t)]} = \frac{\mathrm{Var}(A(t))}{\lambda t}, \quad t \geq 0. \tag{1.1}$$

The second equation follows from the fact that $E[A(t)] = \lambda t$ for stationary point process $A(t)$. The IDC measures the variability over time, independent of the rate λ.

Even though the IDC is a partial characterization of an arrival process, defined directly in terms of the rate and variance-time curve of the arrival process, it characterizes the variability of an arrival process much more completely than the usual variability parameters, such as the variance of a single interarrival time or the lag-1 correlation. Indeed, for a renewal process, the inter-renewal time distribution can be calculated from the rate and the IDC of its stationary (or equilibrium) renewal process, and vise versa. Thus, the $GI/GI/1$ model, involving only renewal processes, is fully specified by the rate and IDC for both the arrival and service stationary counting processes; see [52]. Moreover, Theorem 5 of [50] shows that the new robust queueing algorithm based on indices of dispersion for the general $G/G/1$ queue is asymptotically exact in both light and heavy traffic limits.

1.3 New Theory Supporting the RQNA

Given the strong motivation for working with the IDC, the major challenge is to develop an effective approximation for the IDC of each internal arrival process within the OQN. 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 IDC. 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 [49]. Based on that heavy-traffic limit, in (74) of [49] we developed an approximation of the IDC of a departure process by a convex combination of the IDCs of the arrival and service processes as

$$I_d(t) \approx w_\rho(t) I_a(t) + (1 - w_\rho(t)) I_s(t), \quad t \geq 0, \tag{1.2}$$

where the weight $w_\rho(t)$ has closed-form expression.

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 [51, 53]. The heavy-traffic limits here in the single-bottleneck special case are used in the RQNA algorithm. This limit is exceptionally tractable because they can be expressed in terms of one-dimensional reflected Brownian motion (RBM). The IDC in the heavy-traffic limit can then be calculated in closed-form by applying Corollary 5.1 of [49].

1.4 Organization

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 [46] is asymptotically correct in the HT limit. In §5, we demonstrate how the HT limits in the present paper can be

applied to approximate the IDCs of the stationary flows in a GJN. Finally, we draw conclusions in §6. In the appendix we give an additional literature review on (i) heavy traffic and (ii) the stability of GJN's.

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 [14], Chapter VII of [1], [23] 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 [10], 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 $1 \le i \le K$.

(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 $1 \le i \le K$.

(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 $P(\theta_i^l = e_j) = p_{i,j}$ 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 [12] or Theorem 3.2.1 of [34].

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 [14], which draws on [20], 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 [20]), §4 of [20] 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 $\lambda_0 = (\lambda_{0,1}, \ldots, \lambda_{0,K})$ be the external arrival rate vector and let $\lambda = (\lambda_1, \ldots, \lambda_K)$ 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 $\mathbb{Z}_+^K \times \mathbb{R}_+^{2K}$ for $\mathcal{S}(0)$ such that $\mathcal{S}(t)$ has that same distribution π for all $t \geq 0$. Here $\mathbb{Z}_+$ denote non-negative integers and $\mathbb{R}_+$ denote non-negative real numbers. 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 [14] and [15] that the distribution is unbounded and spread out, see also [14] 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 [23] or Theorem 5.1 of [14] or Theorem 6.2 of [15], which extend earlier work on stability for OQNs in [5], [41] and [22].

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* $t \to \infty$.

If a strong Markov process is Harris recurrent, the existence of a stationary measure (unique up to a constant multiple) is shown in the early [2], which in turn draws on [25]; see also [24]. (More precisely, they assume that the process is a Hunt process). If the measure is finite, it can be normalized to a probability measure and the process is called positive Harris recurrent. It is shown in [14] that $\mathcal{S}$ is positive Harris recurrent, hence the existence and uniqueness of a stationary distribution. [14] assumed Assumption 2.1, Assumption 2.2 and a weaker version of Assumption 2.3: 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 [15], 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. 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, we have*

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

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. 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 [4]. ∎

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

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

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

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

where

$$\mathcal{F}(t) \equiv (A_0(t), A_{\text{int}}(t), A(t), S(t), D(t), D_{\text{ext}}(t)) \tag{2.7}$$

is a vector of cumulative point processes, with the processes defined in §2.1. We refer to $\mathcal{F}$ in (2.7) as the *flows*. We say that a flow is *stationary* if it has stationary increments. We refer to [42] and Chapter 6 of [7] 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

$$S_{i,s}(t) \equiv S_i(B_i(s) + t) - S_i(B_i(s)), \quad \text{for} \quad i = 1, 2, \ldots, K, \tag{2.8}$$

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

$$D_s(t) \equiv D(s + t) - D(s) = (S \odot B)(s + t) - (S \odot B)(s) = (S_s \odot B_s)(t), \quad t \geq 0,$$

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 $s \to \infty$,

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

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 (2.8), 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 $s \to \infty$

$$(Q_s, U_s, V_s, A_{0,s}, S_s) \Rightarrow (Q_e, U_e, V_e, A_{0,e}, S_e)\,. \tag{2.9}$$

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

$$B_{i,s}(t) = \int_s^{s+t} 1_{Q_i(u)>0} du = \int_0^t 1_{Q_{i,s}(u)>0} du = \int_0^t h(Q_{i,s}(u)) du, \quad \text{for} \quad 1 \leq i \leq K. \tag{2.10}$$

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 [7].) 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 (2.10) is a continuous mapping from $\mathcal{D}$ to $\mathcal{C}(\mathbb{R}_+, \mathbb{R})$; see Theorem 11.5.1 in [48]. 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 (2.6) is continuous. Hence, we can extend the convergence as $s \to \infty$ in (2.9) to

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

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 some coordinate of $\mathcal{S}_s$ 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

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

For the types of the jumps, $J_s = (0,i)$, (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. Note that the type is well defined because Assumption 2.3 implies that $P(T_s^a = T_s^d) = 0$.

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)

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

That in turn implies convergence for the associated flow counting processes by applying the inverse map in §13.6 of [48] as stated. For example, we can write

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

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 [10, 11, 14, 38], 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

$$X \equiv Q(t) - (I - P')Y,$$

where Y is defined in (2.6).

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_P be the K-dimensional reflection map with reflection matrix P; e.g., see Chapter 14 of [48].

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

$$X = Q(0) + \tilde{A}_0 + \tilde{\Theta}'\mathbf{1} - (I - P')\tilde{S} + (\lambda_0 - (I - P')\mu)e, \tag{3.2}$$

while the queue length process can be written as

$$Q = X + (I - P')Y = \Psi_{I-P'}(X), \tag{3.3}$$

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

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

where the centered processes are defined in (3.1).

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

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

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 $Q_i = 0$, (3.3) follows from the usual reflection argument.

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

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

or, in matrix notation, by

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

By (3.3), we have

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

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

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

$$
\begin{aligned}
&= \tilde{A}_0 + \tilde{\Theta}'\mathbf{1} + P'\tilde{S} + (\lambda_0 + P'\mu)e \\
&\quad + P'(I - P')^{-1}\left(Q(0) - Q + \tilde{A}_0 + \tilde{\Theta}'\mathbf{1} + \lambda_0 e\right) - P'\tilde{S} - P'\mu e \\
&= P'(I - P')^{-1}(Q(0) - Q) + (I - P')^{-1}\left(\tilde{A}_0 + \tilde{\Theta}'\mathbf{1}\right) + \lambda e. \qquad (3.4)
\end{aligned}
$$

Finally, note that $D = Q(0) + A - Q$. ∎

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.*

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 [38]; 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

$$
\mu_{i,\rho} \equiv \lambda_i/(c_i\rho), \quad 1 \le i \le K, \qquad (3.5)
$$

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 (3.2) is now replaced by the ρ-dependent version in (3.5).

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

$$
\begin{aligned}
A^*_{0,i,\rho}(t) &\equiv (1-\rho)[A_{0,i}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_{0,i}t], \\
A^*_{i,\rho}(t) &\equiv (1-\rho)[A_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i t], \\
S^*_{i,\rho}(t) &\equiv (1-\rho)[S_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\mu_{i,\rho}t], \\
D^*_{i,\rho}(t) &\equiv (1-\rho)[D_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i t], \\
D^*_{\mathrm{ext},i,\rho}(t) &\equiv (1-\rho)[D_{\mathrm{ext},i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i p_{i,0}t],
\end{aligned}
$$

$$
\begin{aligned}
A^*_{i,j,\rho}(t) &\equiv (1-\rho)[A_{i,j,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i p_{i,j} t],\\
\Theta^*_{i,j,\rho}(t) &\equiv (1-\rho)\left[\sum_{l=1}^{\lfloor (1-\rho)^{-2}t\rfloor} \theta^l_{i,j} - p_{i,j}(1-\rho)^{-2}t\right],\\
Q^*_{i,\rho}(t) &\equiv (1-\rho)Q_{i,\rho}((1-\rho)^{-2}t), \text{ for } 1 \le i,j \le K.
\end{aligned} \tag{3.6}
$$

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

$$
\mathcal{F}^*_\rho(t) \equiv (A^*_{0,\rho}(t), A^*_{\text{int},\rho}(t), A^*_\rho(t), S^*_\rho(t), D^*_\rho(t), D^*_{\text{ext},\rho}(t)). \tag{3.7}
$$

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

$$
\hat{P}_{\mathcal{H}} = P_{\mathcal{H}} + P_{\mathcal{H},\mathcal{H}^c}\left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c,\mathcal{H}}. \tag{3.8}
$$

Note that the inverse $\left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1}$ appearing in (3.8) 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

$$
\hat{P}_{\mathcal{H}^c,\mathcal{H}} = \sum_{l=0}^{\infty} (P_{\mathcal{H}^c})^l P_{\mathcal{H}^c,\mathcal{H}} = \left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c,\mathcal{H}}. \tag{3.9}
$$

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

$$
\hat{\lambda}_{0,\mathcal{H}} = \lambda_{0,\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}}\lambda_{0,\mathcal{H}^c} = \lambda_{0,\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}}\left(I_{\mathcal{H}^c} - P'_{\mathcal{H}^c}\right)^{-1}\lambda_{0,\mathcal{H}^c}, \tag{3.10}
$$

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 3.2 (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 $\rho \uparrow 1$,*

$$(Q^*_\rho, Z^*_\rho, \Theta^*_\rho, \Theta^*_{\text{ext},\rho}, \mathcal{F}^*_\rho) \Rightarrow (Q^*, Z^*, \Theta^*, \Theta^*_{\text{ext}}, \mathcal{F}^*), \tag{3.11}$$

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

$$\hat{X}^*_{\mathcal{H}} = Q^*_{\mathcal{H}}(0) + \left(e'_{\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} e'_{\mathcal{H}^c}\right)\left(A^*_0 + (\Theta^*)'\,\mathbf{1}\right) - (I - \hat{P}_{\mathcal{H}})S^*_{\mathcal{H}} - \hat{\lambda}_{0,\mathcal{H}} e \tag{3.12}$$

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 $D_{\text{ext},i}^*$ can be expressed as

$$A_{i,j}^* = p_{i,j} D_i^* + \Theta_{i,j}^* \circ \lambda_i e, \quad \textit{and} \quad D_{\text{ext},i}^* = p_{i,0} D_i^* + \Theta_{i,0}^* \circ \lambda_i e, \quad \textit{for} \quad 1 \le i, j \le K.$$

(vi) The limiting workload process is $Z_i^* = \lambda_i^{-1} Q_i^*$.

Proof. Much of the statement follows from [10, 11] and [8]. First, the HT limit for the state process with an arbitrary subset $\mathcal{H}$ of critically loaded stations follows from [10, 11]. Second, the HT limit for the steady-state queue length follows from [8]. The papers [23] and [8] 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 [8]. 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 (3.6). Thus, the HT-scaled net-input process is

$$X_\rho^* = Q_\rho^*(0) + A_{0,\rho}^* + \left(\tilde{\Theta}_\rho^* \right)' \mathbf{1} - (I - P')\tilde{S}_\rho^* + (\lambda_0 - (I - P')\mu_\rho)(1 - \rho)^{-1} e, \tag{3.13}$$

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

$$Q_{\mathcal{H},\rho}^* = X_{\mathcal{H},\rho}^* + (I - P'_{\mathcal{H},\mathcal{H}})Y_{\mathcal{H},\rho}^* - P'_{\mathcal{H}^c,\mathcal{H}} Y_{\mathcal{H}^c,\rho}^* \tag{3.14}$$

$$Q^*_{\mathcal{H}^c,\rho} = X^*_{\mathcal{H}^c,\rho} + (I - P'_{\mathcal{H}^c,\mathcal{H}^c})Y^*_{\mathcal{H}^c,\rho} - P'_{\mathcal{H},\mathcal{H}^c}Y^*_{\mathcal{H},\rho} \tag{3.15}$$

Solving for $Y^*_{\mathcal{H}^c,\rho}$ in (3.15) and substituting into (3.14), we have

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

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

$$\begin{aligned}\hat{X}^*_{\mathcal{H},\rho} = Q^*_{\mathcal{H},\rho}(0) + \left(e'_{\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\right)\left(A^*_{0,\mathcal{H}^c,\rho} + (\tilde{\Theta}^*_\rho)'\mathbf{1}\right) + (I - \hat{P}'_{\mathcal{H}})\tilde{S}^*_{\mathcal{H},\rho} \\ + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q^*_{\mathcal{H}^c,\rho}(0) - Q^*_{\mathcal{H}^c,\rho}) + \hat{\eta}_\rho(1-\rho)^{-1}e.\end{aligned} \tag{3.17}$$

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

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

Hence

$$\begin{aligned}\hat{\eta}_\rho &\equiv \eta_{\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\eta_{\mathcal{H}^c,\rho} \\ &= \lambda_{0,\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\lambda_{0,\mathcal{H}^c} - (I - \hat{P}'_{\mathcal{H}})\mu_{\mathcal{H},\rho}.\end{aligned} \tag{3.18}$$

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 (3.18), we have

$$\hat{\eta}_\rho = (I - \hat{P}'_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho}). \tag{3.19}$$

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 [48], 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 [11].

For conclusion (ii), we apply [8] 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 [11]. 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

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

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

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

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 [48]. Finally, the associated limits for the workload can be related to the limit for the queue length as indicated in [11]. ∎

Remark 3.3 (Elaboration on the application of [8]) We apply [8], 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 [8]. Comparing (3.6) here with (A5) there, for the bottleneck queues,

the two scaling conventions are connected by setting $n = (1-\rho)^{-2}$, $\tilde{v}_i^n = 0$ and $\tilde{\beta}_i^n = -\lambda_i/\rho$. The stability condition here is then connected to that in [8] 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 [8], 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 (3.19), the rest of the proof is essentially the same. For $Q^*_{\mathcal{H}^c}$, by [10, 11], it is negligible in the sense of Theorem 3.3 of [8]. Theorem 3.4 of [8] 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 [8] remain unchanged.

Remark 3.4 (Functional central limit theorem for the stationary flows) We discuss an important special case of Theorem 3.2 where we set $|\mathcal{H}| = 0$. In this special case, all stations are strictly non-bottleneck, i.e., $\mu_{i,\rho} = \lambda/(\kappa_i \rho)$ where $\kappa_i < 1$ for all i. As $\rho \uparrow 1$, the family of systems converges to a limiting system where the traffic intensity at station i is $\rho_i = \kappa_i$. Hence, the scaling used in (3.6) corresponds to the diffusion scaling used in the usual FCLT. The joint FCLT of the stationary flows can be written as

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

4 Heavy-Traffic Limits with One Bottleneck Queue

In this section we consider the special case in which there is only one bottleneck queue, which is useful for the IDC approximation and the RQNA applications because it is especially tractable, involving one-dimensional RBM instead of multi-dimensional RBM.

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. But then we observe that the case of a single-bottleneck is asymptotically equivalent to that except that the arrival process is generalized to include the immediate feedback associated with flows to all the other non-bottleneck queues.

As a consequence, we show that it is asymptotically correct in HT for a GJN with a single

bottleneck queue to eliminate all feedback prior to analysis. Moreover, we show how to quantify feedback elmination.

4.1 A 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.

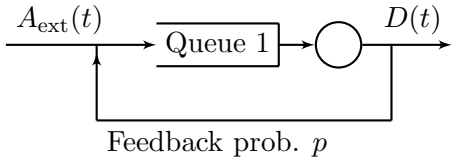

Figure 1: A single-server queue with customer feedback.

Corollary 4.1 (One $GI/GI/1$ queue with feedback) *Under Assumptions in Theorem 3.2, consider a family of single-server queues in stationarity, indexed by ρ. Assume that $\mu_\rho = \lambda/\rho$. Then, as $\rho \uparrow 1$,*

$$(Q_\rho^\star, W_\rho^\star, Z_\rho^\star, \Theta_\rho^\star, \Theta_{\text{ext},\rho}^\star, \mathcal{F}_\rho^\star) \Rightarrow (Q^\star, W^\star, Z^\star, \Theta^\star, \Theta_{\text{ext}}^\star, \mathcal{F}^\star) \quad \textit{in} \quad \mathcal{D}^{11},$$

where $\mathcal{F}_\rho^\star = (A_{0,\rho}^\star, A_\rho^\star, A_{\text{int},\rho}^\star, S_\rho^\star, D_\rho^\star, D_{\text{ext},\rho}^\star)$, $\mathcal{F}^\star = (A_0^\star, A^\star, A_{\text{int}}^\star, S^\star, D^\star, D_{\text{ext}}^\star)$ and:

(i) $A_0^\star = c_{a_0} B_{a_0} \circ \lambda_0 e$ and $S^\star = c_s B_s \circ \lambda e$, where B_{a_0} and B_s are standard Brownian motions. $(\Theta^\star, \Theta_{\text{ext}}^\star)$ is a zero-drift two-dimensional Brownian motion with covariance matrix $\Sigma = (\sigma_{i,j}^2 : 1 \le i,j \le 2)$, where $\sigma_{1,1}^2 = \sigma_{2,2}^2 = p(1-p)\lambda$ and $\sigma_{1,2}^2 = \sigma_{2,1}^2 = -p(1-p)\lambda$, so that

$$\Theta^\star + \Theta_{\text{ext}}^\star = 0.$$

Furthermore, B_{a_0}, B_s and $(\Theta^\star, \Theta_{\text{ext}}^\star)$ are mutually independent.

(ii) The queue length process $Q^\star$ is a stationary one-dimensional RBM

$$Q^\star \equiv \Psi(X^\star),$$

where Ψ is the one-dimensional reflection map and $X^\star$ is a one-dimensional Brownian motion

$$X^\star = Q^\star(0) + A_0^\star + (\Theta^\star - (1-p)S^\star) - \lambda_0 e,$$

where $\lambda_0 = (1-p)\lambda$. *Furthermore,* $Q^\star(0)$ *has unique stationary distribution of the stationary one-dimensional RBM with drift* $-\lambda_0$ *and variance*

$$\lambda_0 c_x^2 \equiv \lambda_0 \left(c_a^2 + p + (1-p)c_s^2\right),$$

so an exponential distribution with mean $c_x^2/2$.

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

$$A^\star = \frac{1}{1-p}\left(A_0^\star + \Theta^\star\right) + \frac{p}{1-p}\left(Q^\star(0) - Q^\star\right).$$

(iv) The stationary total departure process $D^\star$ *is specified as*

$$\begin{aligned} D^\star &= Q^\star(0) + A^\star - Q^\star \\ &= \frac{1}{1-p}\left(Q^\star(0) - Q^\star + A_0^\star + (\Theta^\star)'\,\mathbf{1}\right) \end{aligned} \tag{4.1}$$

(v) The internal arrival flow $A_{\text{int}}^\star$ *can be expressed as*

$$A_{\text{int}}^\star = pD^\star + \Theta^\star$$

and the external departure flow can be expressed as

$$D_{\text{ext}}^\star = (1-p)D^\star + \Theta_{\text{ext}}^\star = A_0^\star + Q^\star(0) - Q^\star.$$

(vi) $Z^\star = \lambda^{-1}Q^\star$ *and* $W^\star = Z^\star \circ \lambda e$.

As observed in Section III of [46], 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 $\tilde{S}^\star \equiv \tilde{c}_s^2 \tilde{B}_s \circ \lambda_0 e$. We now claim that $\tilde{S}^\star \stackrel{dist.}{=} \Theta^\star - (1-p)S^\star$. To this end, note that $\Theta^\star = \sqrt{p(1-p)}B_\Theta \circ \lambda e$ and $S^\star = c_s B_s \circ \lambda e$, where B_Θ, B_s are independent standard Brownian motions (zero drift and unit variance) and $\lambda_0 = (1-p)\lambda$. Hence, from part (ii) of Corollary 4.1, we have

$$X^\star \stackrel{dist.}{=} Q^\star(0) + A_0^\star + \tilde{S}^\star - \lambda_0 e. \tag{4.2}$$

Let $\tilde{Q}^\star, \tilde{Z}^\star, \tilde{W}^\star$ denote the HT limit of the queue length process, the workload process and the waiting time 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 (4.2) is exactly the HT limit of the net-input process of a single-server queue so that $\tilde{Q}^\star \stackrel{dist.}{=} Q^\star$. Hence, we have

$$\begin{aligned}
\tilde{Z}^\star \equiv \lambda_0^{-1}\tilde{Q}^\star &\stackrel{dist.}{=} (1-p)^{-1}\lambda^{-1}Q^\star \equiv (1-p)^{-1}Z^\star, \quad \text{and} \\
\tilde{W}^\star \equiv \tilde{Z}^\star \circ \lambda_0 e &\stackrel{dist.}{=} (1-p)^{-1}Z^\star \circ \lambda_0 e \equiv (1-p)^{-1}W^\star \circ (1-p)e.
\end{aligned}$$

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 Corollary 4.1, 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*

$$(Q^\star, Z^\star, W^\star, D_{\text{ext}}^\star) \stackrel{dist.}{=} (\tilde{Q}^\star, (1-p)\tilde{Z}^\star, (1-p)\tilde{W}^\star \circ (1-p)^{-1}e, \tilde{D}_{\text{ext}}^\star).$$

4.2 Networks with One Bottleneck Queue

We now consider the more general special case in which $K \geq 1$ but $|\mathcal{H}| = 1$. Without loss of generality, let $\mathcal{H} = \{h\}$, so that station h is the only bottleneck station. Then Theorem 3.2 can be restated as

Corollary 4.2 (Network with one bottleneck queue) *Under Assumption 2.1-2.2, consider a series of GJNs in stationarity, indexed by ρ. Assume that $\mu_{i,\rho} = \lambda_i/(\kappa_i \rho)$ for $1 \leq i \leq K$ and set $c_h = 1$ and $\kappa_i < 1$ for all $i \neq h$. Then, we have*

$$(Q_\rho^\star, W_\rho^\star, Z_\rho^\star, \Theta_\rho^\star, \Theta_{\text{ext},\rho}^\star, \mathcal{F}_\rho^\star) \Rightarrow (Q^\star, W^\star, Z^\star, \Theta^\star, \Theta_{\text{ext}}^\star, \mathcal{F}^\star)$$

as $\rho \uparrow 1$ *in* $\mathcal{D}^{9K+2K^2}$*, where:*

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

(ii) The queue length process $Q^\star$ *consists of two parts.* $Q^\star_i \equiv 0$ *for* $i \ne h$ *and* $Q^\star_h$ *is a stationary one-dimensional RBM*

$$Q^\star_h \equiv \Psi\left(\hat{X}^\star_h\right),$$

where Ψ *is the one-dimensional refelction map and* $\hat{X}^\star_h$ *is the net-input process defined as*

$$\hat{X}^\star_h = Q^\star_h(0) + \left(e'_h + \hat{P}'_{\mathcal{H}^c,h} e'_{\mathcal{H}^c}\right)\left(A^\star_0 + (\Theta^\star)'\mathbf{1}\right) - (1-\hat{P}_h)S^\star_h - \hat{\lambda}_{0,h} e, \tag{4.3}$$

where $e_{\mathcal{A}}$ *collects columns in the* K*-dimensional identity matrix* I *that corresponds to index set* $\mathcal{A}$*;* $\hat{P}_h$*,* $\hat{P}_{\mathcal{H}^c,h}$ *and* $\hat{\lambda}_{0,h}$ *are defined in (3.8), (3.9) and (3.10) with* $\mathcal{H} = \{h\}$*, respectively. Furthermore,* $Q^\star_h(0)$ *has unique stationary distribution of the stationary RBM.*

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

$$A^\star = (I-P')^{-1}\left(A^\star_0 + (\Theta^\star)'\mathbf{1}\right) + P'(I-P')^{-1} e_h \left(Q^\star_h(0) - Q^\star_h\right).$$

(iv) The stationary departure process $D^\star$ *is specified as*

$$D^\star = Q^\star(0) + A^\star - Q^\star = (I-P')^{-1}\left(Q^\star(0) - Q^\star + A^\star_0 + (\Theta^\star)'\mathbf{1}\right).$$

In particular,

$$D^\star_{\mathcal{H}^c} = Q^\star_{\mathcal{H}^c} + A^\star_{\mathcal{H}^c} - Q^\star_{\mathcal{H}^c}(0) = A^\star_{\mathcal{H}^c}.$$

(v) The internal arrival flow $A^\star_{i,j}$ *can be expressed as*

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

and the external departure flow can be expressed as

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

(vi) $Z_i^\star = \lambda_i^{-1} Q_i^\star$ *and* $W_i^\star = Z_i^\star \circ \lambda_i e$.

We conclude this section by observing that in 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 also show that feedback elimination is asymptotically correct for networks with one bottleneck.

We start with a convenient representation of the HT limit of the bottleneck queue. Let $\hat{p}_{i,h}$ be the (i,h)-th component of $\hat{P}_{\mathcal{H}^c,\mathcal{H}}$ in *(3.9)* and recall that $\hat{p} \equiv \hat{P}_h$ is the feedback probability defined in Remark 3.2.

Theorem 4.2 *The HT limit* $\hat{X}_h^\star$ *in* (4.3) *can be expressed as the following one-dimensional Brownian motion*

$$\hat{X}_h^\star = Q_h^\star(0) + \hat{A}^\star + \left(\hat{\Theta}_S^\star - (1-\hat{p})S_h^\star\right) + \hat{\lambda}_{0,h} e, \tag{4.4}$$

where

$$\hat{A}^\star = A_{0,h}^\star + \sum_{i \in \mathcal{H}^c} \left(\hat{p}_{i,h} A_{0,i}^\star + \hat{\Theta}_{i,h}^\star\right), \tag{4.5}$$

and

$$\begin{aligned} \hat{\Theta}_{i,h}^\star &= \sqrt{\hat{p}_{i,h}(1-\hat{p}_{i,h})} B_{\hat{\Theta}_{i,h}} \circ \lambda_{0,i} e, \\ \hat{\Theta}_S^\star &= \sqrt{\hat{p}(1-\hat{p})} B_{\hat{\Theta}_S} \circ \lambda_i e, \end{aligned} \tag{4.6}$$

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 $A_0^\star$ and $S_h^\star$ 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

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

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, comparing Theorem 4.2 with Corollary 4.1, we conclude that both the reduced network and the original bottleneck queue is asymptotically equivalent to 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 $\hat{S}^\star \equiv \hat{\Theta}^\star_S - (1-\hat{p})S^\star_h$, where $\Theta^\star_S$ is defined in (4.6), just as discussed in Section 4. This matches exactly with the "service" part in (4.4) of Theorem 4.2. Hence, we have the following theorem, extending 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*

$$(Q^\star, Z^\star, W^\star, D^\star_{\rm ext}) \stackrel{dist.}{=} (\hat{Q}^\star, (1-p)\hat{Z}^\star, (1-p)\hat{W}^\star \circ (1-p)^{-1}e, \hat{D}^\star_{\rm ext}).$$

5 Approximation of the IDC

In this section, we demonstrate how the HT limits in the present paper can be applied to approximate the IDCs of the stationary flows in a GJN, where the IDC is defined in (1.1). In particular, we focus on two simple examples, one for the superposition operation and one for the splitting operation.

5.1 Dependent Superposition: Splitting and Re-Combining

Dependence among flows are ubiquitous in GJNs. Even in a feed-forward network, there can be dependence among the arrival processes being superposed at one of the queues in the network. That is illustrated by an example in Figure 2 where an arrival process is first split into two streams according to Markovian routing and sent to separate queues, and then the two departure processes are recombined to enter a third queue. We aim to approximate the IDC of the superposition of the two stationary departure processes $A_3(t) \equiv D_1(t) + D_2(t)$. To do so, we establish the HT limit for the superposition arrival process at the third queue.

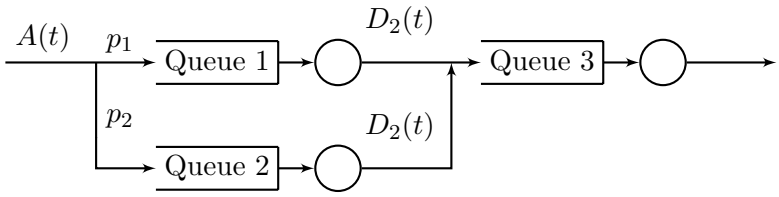

Figure 2: A re-combining after splitting example.

Without loss of generality, assume that the traffic intensity ρ_1 at the first queue is larger than ρ_2 at the second queue. We then consider a family of systems indexed by ρ, where the traffic intensity at queue 1 is $\rho_1 = \rho$, which we will bring to heavy traffic, and the traffic intensity at queue 2 is fixed at $\rho_2 \in [0, 1)$. Let $A_{i,\rho}$, $S_{i,\rho}$ and $Q_{i,\rho}$ denote the arrival process, the (uninterrupted) service renewal processes and the queue length process at Queue i in the ρ-th system, respectively.

Corollary 5.1 (Heavy-traffic limit for Splitting and Recombining) *Consider the system depicted in Figure 2. Assume that the external arrival process is renewal with rate λ and scv c_a^2, the service times at queue 1 are i.i.d. with rate $p_1\lambda/\rho$ and scv $c_{s_1}^2$; the service times at queue 2 are i.i.d. with rate $p_2\lambda/\rho_2$ for $0 \le \rho_2 < 1$ and scv $c_{s_2}^2$. Then*

$$(A_\rho^*, A_{1,\rho}^*, A_{2,\rho}^*, S_{1,\rho}^*, S_{2,\rho}^*, Q_{1,\rho}^*, Q_{2,\rho}^*, D_{1,\rho}^*, D_{2,\rho}^*, \Theta_{1,\rho}^*, \Theta_{2,\rho}^*)$$
$$\Rightarrow (A^*, A_1^*, A_2^*, S_1^*, S_2^*, Q_1^*, Q_2^*, D_1^*, D_2^*, \Theta_1^*, \Theta_2^*) \quad in \quad \mathcal{D}^{11} \quad as \quad \rho \to 1,$$

where

$$
\begin{aligned}
&A^* \equiv c_a B_a \circ \lambda e,\\
&A_i^* \equiv p_i c_a B_a \circ \lambda e + \Theta_i^*, \quad \textit{for} \quad i = 1, 2,\\
&S_1^* \equiv c_{s_1} B_{s_1} \circ p_1 \lambda e,\\
&S_2^* \equiv c_{s_2} B_{s_2} \circ p_2 \lambda e / \rho_2,\\
&Q_1^* \equiv \psi(Q_1^*(0) + p_1 c_a B_a \circ \lambda e + \Theta_1^* - c_{s_1} B_{s_1} \circ p_1 \lambda e - p_1 \lambda e)\\
&Q_2^* \equiv 0,\\
&D_1^* \equiv p_1 c_a B_a \circ \lambda e + \Theta_1^* + Q_1^*(0) - Q_1^*,\\
&D_2^* \equiv p_2 c_a B_a \circ \lambda e + \Theta_2^*,
\end{aligned} \tag{5.1}
$$

with ψ being the one-dimensional reflection mapping and (Θ_1^*, Θ_2^*) being a zero-drift two-dimensional Brownian motion with covariance matrix $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{J \times J}$, where $\sigma_{ii}^2 = p_i(1-p_i)\lambda$ and $\sigma_{ij}^2 = -p_i p_j \lambda$ for $i \neq j$.

To approximate the IDC of the total arrival process at queue 3, we write

$$
\begin{aligned}
I_{a,3,\rho}(t) &\equiv \frac{\mathrm{Var}(A_{3,\rho}(t))}{E[A_{3,\rho}(t)]} = \frac{\mathrm{Var}\left(D_{1,\rho}(t) + D_{2,\rho}(t)\right)}{E[A_{3,\rho}(t)]}\\
&= \frac{\mathrm{Var}\left(D_{1,\rho}(t)\right)}{E[A_{3,\rho}(t)]} + \frac{\mathrm{Var}\left(D_{2,\rho}(t)\right)}{E[A_{3,\rho}(t)]} + \mathrm{cov}\left(D_{1,\rho}(t), D_{2,\rho}(t)\right) / E[A_{3,\rho}(t)]\\
&= p_1 I_{d,1,\rho}(t) + p_2 I_{d,2,\rho}(t) + \beta_\rho(t),
\end{aligned}
$$

where

$$
\beta_\rho(t) \equiv \mathrm{cov}\left(D_{1,\rho}(t), D_{2,\rho}(t)\right) / E[A_{3,\rho}(t)]. \tag{5.2}
$$

In general, exact characterization of β_ρ is not readily available. We propose the following approximation

$$
\begin{aligned}
\beta_\rho(t) &\approx 2\mathrm{cov}\left(D_1^*((1-\rho)^2 t), D_2^*((1-\rho)^2 t)\right) / (\lambda(1-\rho)^2 t)\\
&= 2p_1(1-p_1)(c_{a_0}^2 - 1) w^*((1-\rho)^2 p_1 \lambda t / c_{x_1}^2))
\end{aligned} \tag{5.3}
$$

with D_1^* and D_2^* being the diffusion limit in (5.1).

To justify the approximation (5.3), let $\beta_\rho^*(t) = \beta_\rho\left((1-\rho)^{-2} t\right)$ be the HT-scaled correction term. Corollary 5.1 implies the following limit.

Corollary 5.2 *Under the assumption in Theorem 5.1 and the exchange of limit assumptions, we have*

$$
\beta_\rho^* \Rightarrow 2p_1(1-p_1)(c_{a_0}^2 - 1) w^*\left(p_1 \lambda t / c_{x_1}^2\right). \tag{5.4}
$$

Proof Note that Corollary 5.1 implies that

$$\begin{aligned}\mathrm{cov}(D_{1,\rho}(t), D_{1,\rho}(t)) &= \mathrm{cov}\left((1-\rho_1)^{-1}D^*_{1,\rho}((1-\rho_1)^2 t), (1-\rho_1)^{-1}D^*_{2,\rho}((1-\rho_1)^2 t)\right) \\ &\Rightarrow (1-\rho_1)^{-2}\mathrm{cov}(D^*_1((1-\rho_1)^2 t), D^*_2((1-\rho_1)^2 t)),\end{aligned}$$

as $\rho \uparrow 1$.

On the other hand, by applying Corollary 5.1 of [49], we have

$$\begin{aligned}\mathrm{cov}\left(D^*_1(t), D^*_2(t)\right) &= \mathrm{cov}(A^*_1(t), A^*_2(t)) - \mathrm{cov}(Q^*_1(t), A^*_2(t)) \\ &= p_1(1-p_1)(c^2_{a_0}-1)\lambda t - \mathrm{cov}(Q^*_1(t)), A^*_2(t)) \\ &= p_1(1-p_1)(c^2_{a_0}-1)\lambda t w^*\left(p_1\lambda t/c^2_{x_1}\right),\end{aligned}$$

where $c^2_{x_1} = c^2_{a_1} + c^2_s$, $c^2_{a_1} = p_1 c^2_a + (1-p_1)$ and w^* is the weight function defined in (28) of [49]. The limit then follows. ∎

We demonstrate the performance of the approximation by making simulation comparisons in Example 5.1.

Example 5.1 (splitting and recombining) Consider the queueing system in Figure 2 with rate-1 hyperexponential ($H_2(4)$) external arrival process and $c^2_a = 4$, $p_1 = 0.25$, $p_2 = 0.75$ and i.i.d. Erlang (E_2) service times with $c^2_{s_i} = 0.5$. Figure 3 shows the results for two cases involving different traffic intensities: (i) $\rho_1 = \rho_2 = 0.7$ (left); and (ii) $\rho_1 = 0.8$ and $\rho_2 = 0.9$ (right). In each plot, we display, in solid lines, the IDC $I_{a,3}$ of the total arrival process at queue 3, the modified IDC's $p_i I_{d,i}$ of the departure processes from queue i, the simulated correction term β_ρ defined in (5.2). For approximations, we display, in broken lines, the approximated correction terms as in (5.3) and the approximated IDC using (5.3). Figure 3 shows remarkable agreement of the approximation and the simulation estimate.

5.2 Dependent Splitting: One Queue with Immediate Feedback

Consider the single-server queue with immediate customer feedback as in §4.1. This introduce dependence between the splitting decision process and the arrival process.

For the splitting operation, suppose that the splitting decision is independent of the departure process, then by the conditional variance formula, we have

$$\mathrm{Var}(A_{\mathrm{int}}(t)) = p^2\mathrm{Var}(D(t)) + p(1-p)\lambda t,$$

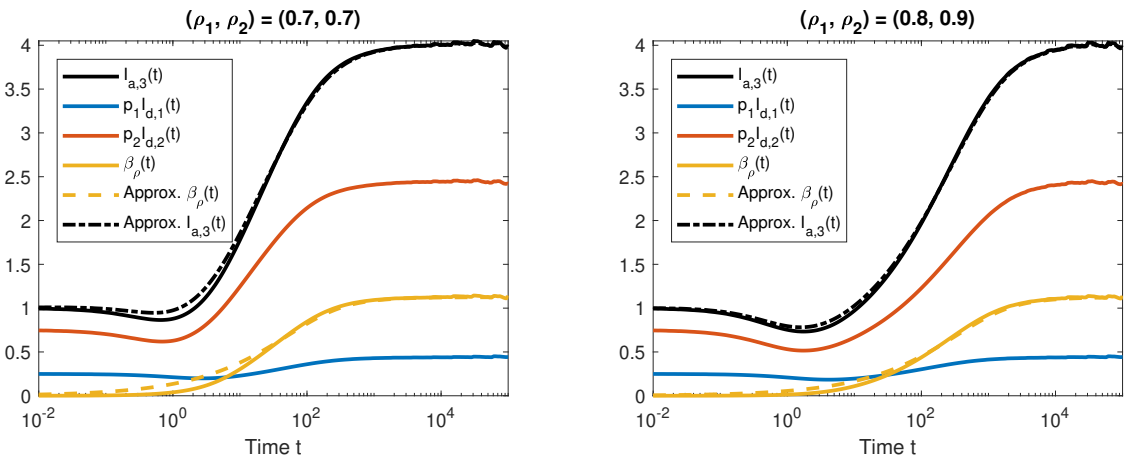

Figure 3: Two examples in Example 5.1.

or equivalently, since $E[D(t)] = \lambda t$ and $E[A_{\text{int}}(t)] = p\lambda t = pE[D(t)]$,

$$I_{a,\text{int}}(t) = pI_d(t) + (1-p).$$

To address the impact of dependence on the IDC after the splitting operation, we propose to consider the correction term $\alpha(t)$ is defined as

$$\alpha(t) \equiv I_{a,\text{int}}(t) - pI_d(t) - (1-p),$$

so that

$$I_{a,\text{int}}(t) = pI_d(t) + (1-p) + \alpha(t), \tag{5.5}$$

We propose to approximate the correction term $\alpha(t)$ by

$$\alpha(t) \approx \alpha^*((1-\rho)^2 t) \tag{5.6}$$

with

$$\alpha^*(t) \equiv 2\text{cov}(pD^*(t), \Theta^*(\lambda t))/p\lambda t = 2pw^*(t/c_x^2),$$

where $c_{x_1}^2 = c_a^2 + c_s^2$, $c_a^2 = \frac{1}{1-p}c_{a_0}^2 + \frac{p}{1-p}$, w^* is the weight function defined in (28) of [49] and the explicit expression is derived using Corollary 5.1 of [49].

The approximation (5.6) is supported by the following corollary. Define the HT-scaled correction term $\alpha_\rho^*(t) \equiv \alpha((1-\rho)^{-2}t)$.

Corollary 5.3 *Under the assumptions in Theorem 4.1 plus the uniform integrability conditions, we have* $\alpha_\rho^*(t) \Rightarrow \alpha^*(t)$ *as* $\rho \uparrow 1$.

Proof By the definitions of the correction term and HT-scaled processes, we write

$$\begin{aligned}
\alpha^*_\rho(t) &= \alpha((1-\rho)^{-2}t) \\
&= I_{a,\mathrm{int}}((1-\rho)^{-2}t) - pI_d((1-\rho)^{-2}t) - (1-p) \\
&= \frac{\mathrm{Var}((1-\rho)A_{\mathrm{int}}((1-\rho)^{-2}t))}{p\lambda_i t} - p\frac{\mathrm{Var}((1-\rho)D((1-\rho)^{-2}t))}{\lambda t} - (1-p) \\
&= \frac{\mathrm{Var}(A^*_{\mathrm{int},\rho}(t))}{p\lambda t} - p\frac{\mathrm{Var}(D^*_\rho(t))}{\lambda t} - (1-p) \\
&\Rightarrow \frac{\mathrm{Var}(A^*_{\mathrm{int}}(t))}{p\lambda_i t} - p\frac{\mathrm{Var}(D^*(t))}{\lambda t} - (1-p) = \alpha^*(t). \quad \blacksquare
\end{aligned}$$

Finally, we also have dependent superposition in this example. Similar to §5.1, we have

$$I_{a,\rho}(t) \approx \frac{1}{1-p} I_{a,0,\rho}(t) + \frac{p}{1-p} I_{a,\mathrm{int},\rho}(t) + \beta_\rho(t) \tag{5.7}$$

with

$$\begin{aligned}
\beta_\rho(t) &\equiv 2\mathrm{cov}(A^*_0((1-\rho)^2 t), A^*_{\mathrm{int}}((1-\rho)^2 t))/(\lambda(1-\rho)^2 t) \\
&= 2pc^2_{a_0} w^*((1-\rho)^2/c^2_x),
\end{aligned} \tag{5.8}$$

where again $c^2_x = c^2_a + c^2_s$ and $c^2_a = \frac{1}{1-p}c^2_{a_0} + \frac{p}{1-p}$.

We demonstrate the performance of the approximation by making simulation comparisons in Example 5.2.

Example 5.2 (immediate feedback) Figure 4 compares the performance of the IDC approximation to simulations for the $E_2/H_2(4)/1$ single-server queue with feedback model, having service scv $c^2_s = 4$. The plot on the left focuses on the feedback flow $A_{\mathrm{int}}(t)$, while the plot on the right focuses on the superposition arrival process $A(t)$. Again, the approximation matches simulation remarkably well.

6 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 (2.7), 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 (3.6)

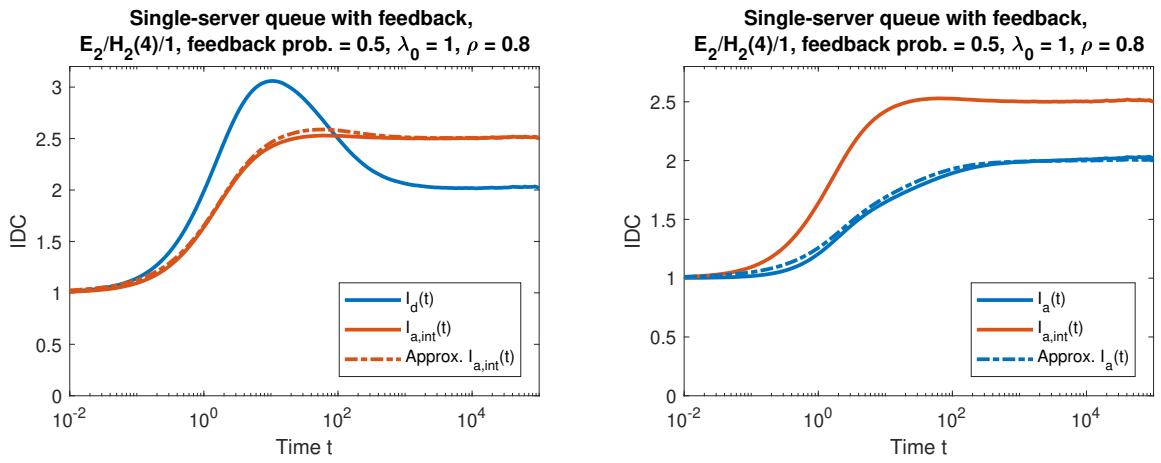

Figure 4: Left plot shows the dependent splitting in a single-server queue with feedback example. Model parameters are described in the title. The simulation estimation of the IDC of the feedback flow is contrasted to the IDC approximation *(5.5)* with correction term *(5.6)* in dotted-and-dashed lines. Right plot displays the dependent superposition. The simulation estimation of the IDC of the total arrival process is contrasted to the IDC approximation *(5.7)* with correction term *(5.8)* in dotted-and-dashed lines.

and (3.7). We then obtained explicit results for the special case in which zero or one station is critically loaded in §4. Finally, we experimentally confirmed the theorems and illustrated how they can be applied to RQNA by considering two examples involving (i) dependent superposition and (ii) dependent splitting in §5.

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 external arrival processes, which requires generalizing the key supporting theorems in [8, 23]. 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, Hanqin Zhang 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] F. Baccelli and P. Bremaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*. Springer, New York, second edition, 2003.

[4] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1999.

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

[6] A. Braverman, J. G. Dai, and M. Miyazawa. Heavy traffic approximation for the stationary distribution of a generalized Jackson network: the BAR approach. *Stochastic Systems*, 7(1):143–196, 2017.

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

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

[9] C. Chang, J. Thomas, and S. Kiang. On the stability of open networks: a unified approach by stochastic dominance. *Queueing Systems*, 15(1-4):239–260, 1994.

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

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

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

[13] D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events*. Methuen, London, 1966.

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

[15] 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.

[16] 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.

[17] D. J. Daley. Queueing output processes. *Adv. Appl. Prob.*, 8(2):395–415, 1976.

[18] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Elementory Theory and Methods*, volume I. Springer, Oxford, U. K., second edition, 2008.

[19] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: General Theory and Structure*, volume II. Springer, Oxford, U. K., second edition, 2008.

[20] 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.

[21] R. L. Disney and D. Konig. Queueing networks: a survey of their random processes. *SIAM Review*, 27 (3):335–403, 1985.

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

[23] 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.

[24] R. K. Getoor. Transience and recurrence of Markov processes. In *Séminaire de Probabilités XIV 1978/79*, pages 397–409. Springer, 1979.

[25] 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.

[26] J. M. Harrison. The heavy traffic approximation for single server queues in series. *Journal of Applied Probability*, 10(3):613–629, 1973.

[27] J. M. Harrison. The diffusion approximation for tandem queues in heavy traffic. *Advances in Applied Probability*, 10(4):886–905, 1978.

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

[29] J. M. Harrison and M. I. Reiman. Reflected Brownian motion on an orthant. *The Annals of Probability*, pages 302–308, 1981.

[30] J. M. Harrison and R. J. Williams. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics: An International Journal of Probability and Stochastic Processes*, 22(2):77–115, 1987.

[31] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic, I. *Advances in Applied Probability*, 2(1):150–177, 1970.

[32] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic, II: Sequences, networks and batches. *Advances in Applied Probability*, 2(2):355–369, 1970.

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

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

[35] P. Konstantopoulos and J. Walrand. Stationary and stability of fork-join networks. *Journal of Applied Probability*, 26(3):604–614, 1989.

[36] B. Melamed. On Poisson traffic processes in discrete-state Markovian systems with applications to queueing theory. *Advances in Applied Probability*, 11(1):218–239, 1979.

[37] S. P. Meyn and S. Down. Stability of generalized Jackson networks. *The Annals of Applied Probability*, pages 124–148, 1994.

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

[39] M. I. Reiman. Asymptotically exact decomposition approximations for open queueing networks. *Operations research letters*, 9(6):363–370, 1990.

[40] K. Sigman. Queues as Harris recurrent Markov chains. *Queueing Systems*, 3(2):179–198, 1988.

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

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

[43] A. L. Stolyar. On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. *Markov Processes and Related Fields*, 1(4):491–512, 1995.

[44] J. Walrand. Poisson flows in single-class open networks of quasireversible queues. *Soch. Proc. Appl.*, 13:292–303, 1982.

[45] J. Walrand. *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

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

[47] W. Whitt. Queues with superposition arrival processes in heavy traffic. *Stochastic Processes and Their Applications*, 21:81–91, 1985.

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

[49] 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.

[50] 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.

[51] 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.

[52] W. Whitt and W. You. The advantage of indices of dispersion in queueing approximations. Operations Research Letters, forthcoming. Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2019.

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

A Appendix: Additional Literature Review

A.1 Heavy Traffic

A major source of approximations for GJNs has been heavy-traffic (HT) limits, first for feed-forward networks in [31, 32] and [26, 27]. As indicated in §IV.3 of [46], the approximation for superposition processes there draws on the HT limit in [47].

New approximations for GJNs have been based on Reiman [38]. In [38] the HT limit of the vector queue length process is shown to be a reflected Brownian motion (RBM) on the nonnegative orthant. The concept of RBM is first introduced in the queueing settings in [27] and studied in detail in [29]. In [10, 11] HT limits were extended to models with strict bottlenecks ($\rho_i > 1$) and non-bottleneck stations ($\rho_i < 1$) as well as the usual critically loaded stations ($\rho_i = 1$). (We do not consider strict bottlnecks here.)

These heavy-traffic limits served as a theoretical basis for the QNET and SBD approximations in [28], [39], and [16]. Theoretical justification for the approximation of the steady-state performance in the GJN by the steady-state performance of the limiting RBM was established by [23] and [8] when they justified interchanging the limits $t \to \infty$ and $\rho \to 1$. Recently direct heavy-traffic limits have been established for the stationary distributions by [6].

So far, the heavy-traffic literature has focused on the queue length, busy time, waiting time, workload and the sojourn time processes. However, little is known beyond the initial results in [31, 32] regarding the HT limits of the arrival flows and departure flows.

A.2 Stability of GJNs

There is a substantial literature on the existence of a proper steady state and the convergence to it; This is referred to as the stability of an open queueing network.

The standard approach has been to focus on the Markov process consisting of the queue length process and the residual interarrival times and service times in the GJN. Early study of such Markov processes includes [5], which considered a slightly different open queueing network (a station is picked to act as both the source and the sink) and proved the convergence of the distribution of the queue length process to a stationary distribution. The stability of a network without feedback is considered in [35]. Sigman [40, 41] showed that the general open queueing network is Harris recurrent and the distribution of the Markov process converges if and only if the interarrival distribution is spread-out; see also [9] for a different approach to stability via stochastic dominance. However, [41] and [9] assumed that there is a single external arrival process that is split to create

arrivals to the individual queues. Harris recurrence for the general case was established by Dai [14], but under the extra condition that each interarrival-time distribution is unbounded above. [14] was primarily concerned with the harder (and interesting) multi-class model, which was also studied in [15, 43]. (We do not consider the multi-class model here.) In [37] the stronger convergence in mean for queue length process and total workload process was established under slightly more restrictive conditions. In [30], a Brownian model for the OQN is considered and the stability result is established.

The existing literature is quite extensive, but it has focused on the stability of the queue length, instead of the flows in the open queueing network. As far as we know, we are the first to consider the stability of the flows.