Stochastic Systems

HEAVY-TRAFFIC LIMITS FOR STATIONARY NETWORK FLOWS

By Ward Whitt[*], and Wei You[*]

November 28, 2018

We establish heavy-traffic limits for the stationary flows in generalized Jackson networks, allowing an arbitrary subset of the queues to be critically loaded. The flows are the processes counting customers flowing from one queue to another or out of the network. 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 leads to accurate approximation of the index of dispersion for counts (IDC), which plays a crucial role in our robust queueing network analyzer (RQNA) for approximating the steady-state performance of a non-Markovian open queueing network.

1. Introduction. The purpose of this paper is to develop a better understanding of the stationary flows in a non-Markov open queueing network (OQN), i.e., the departure processes, flows from one queue to another, superpositions of such processes and thus the internal arrival processes. In many queueing systems, there may be as much interest in the departure process as in the familiar measures of congestion such as queue lengths and waiting times, because the departure process may represent the flow of completed work over time. Thus, it is natural to be interested in the stochastic variability of the flows as well as the relatively tractable rate.

The flows are special stochastic point processes, for which there is a well-developed general theory, as in [19, 20]. There also is a substantial literature on the general structure of stationary point processes in queueing systems, as in Chapter 1 of [3] and [41], 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, for which there is a substantial theory, as in Ch. 4 of [44], 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

[*]Department of Industrial Engineering and Operations Research, Columbia University

Keywords and phrases: generalized Jackson network, heavy traffic, stationary processes, index of dispersion, queueing networks, robust queueing

are Poisson if and only if they are not part of a loop; see [35, 43]. For non-Markov open networks, the flows are even more complicated. As discussed in [18, 21] and references there, the stationary departure process from a $GI/GI/1$ queue is Poisson if and only if the queue is an $M/M/1$ queue.

In this paper we contribute by establishing heavy-traffic (HT) limits for the stationary flows in an OQN of single-server queues. This paper is a sequel to [51], which established a HT limit from the stationary departure process from the $GI/GI/1$ queue. That evidently was the first HT limit for a stationary flow in an OQN. In particular, here 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 often called a *generalized Jackson network* (GJN), because it generalizes the Markovian OQN analyzed by Jackson [31] in which all the interarrival times and service times have exponential distributions. The 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.

In addition to contributing to a better understanding of flows in GJNs, which is important in its own right, we apply the HT limits here in [49] to develop a new approximation method for non-Markov OQNs, which is important because relatively little is known about the exact steady-state performance of a GJN. A GJN is relatively easy to simulate, but there are few analytical formulas showing the performance impact of key parameters. Early analytical approximations were based on the parametric-decomposition method as in [34] and [45], which acts as if the product form still holds, with the performance of each queue approximated by an appropriate function of the exact arrival rate (the same as for a Jackson network) and appropriate variability parameters. A fast algorithm is produced if the variability parameters can be obtained as the unique solution to a set of linear equations, just like the arrival rates from the traffic rate equations, as in §IV.2 of [45].

An alternative way to develop approximations for GJNs is to apply heavy-traffic limits based on Reiman [37]. That has led to the QNET and sequential bottleneck decomposition (SBD) approximations in [26], [38], and [16]. These methods require calculating the steady-state distribution of multidimensional reflected Brownian motion, exploiting [17].

Recently, we began studying an alternative non-parametric stochastic modeling approach, where the flows are partially characterized by their rates and indices of dispersion for counts (IDC). A robust queueing technique is then applied to convert the IDC characterization of the GJN into approximations for the steady-state performances, see [50] and the references therein. 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

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

The IDC measures the variability over time, independent of the rate λ.

Even though the IDC is only a partial characterization of the arrival process involving the variance, 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. In [48], several advantages of IDC in queueing approximation are exposed: (i) the IDC, together with the rate, fully characterizes a renewal process; (ii) combining with the RQ algorithm in [50], it produces accurate approximations for performance measures in GJNs; and (iii) the approximation is much more adaptive to complex distributions than traditional parametric methods.

The IDCs of external arrival processes are readily available: (i) first, explicitly for many stochastic arrival processes, e.g., see §III.G. of [22], which draws on [12]; (ii) second, can be calculated by numerical inversion for renewal processes, as indicated in Chapter 13 of [1]; and (iii) finally, can be estimated from simulation or system data otherwise.

The challenge lies in the calculation or estimation of the IDC of the internal arrival processes. For approximation of the GJNs, we developed a full robust queueing network analyzer (RQNA) for GJNs based on the IDC; see §6 of [50], §6 of [51] and [49]. The effectiveness of this IDC-based RQNA largely depends on our ability to approximate the IDC of each of the internal arrival processes at the queues, which combines flows from other queues with its own external arrivals.

In (74) of [51] 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 $w_\rho(t)$ is based on a HT limit for the stationary departure processes, which shows that it is asymptotically correct for the $GI/GI/1$ model in the HT limit.

The present paper contributes to IDC approximation by establishing HT limits for all the stationary flows in a GJN, allowing any subset of the stations to be bottleneck stations (critically loaded in the limit). The HT limits are especially tractable in the case of a single bottleneck station, because they can be expressed in terms of one-dimensional reflected Brownian motion (RBM). The IDC in HT limit can be calculated in closed-form by applying Corollary 5.1 of [51] and pre-limit approximation can be dereived, see §4.3 of [49]. The limits in this single-bottleneck special case are used in RQNA. The numerical examples in §7 of [49] show that the IDC-based RQNA is quite effective, comparable to the highly successful SBD for the examples considered in [16], but without analyzing multidimensional RBM.

Our results in this paper extend the heavy-traffic limit of the stationary departure process in the $GI/GI/1$ model in [51]. As before, we rely heavily on the justification for interchanging the limits $t \to \infty$ and $\rho \to 1$ in a GJN provided by Gamarnik and Zeevi [23] and Budharija and Lee [7]. 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 [9, 10]. Our main contributions here are the HT limits for the stationary flows.

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, drawing on Sigman [39, 40] and Dai [14]; see Ch. VII of Asmussen [2].

1.1. *Literature Review.*

1.1.1. *Heavy Traffic.* A major source of approximations for GJNs has been heavy-traffic (HT) limits, first for feed-forward networks in [29, 30] and [24, 25]. As indicated in §IV.3 of [45], the approximation for superposition processes there draws on the HT limit in [46].

New approximations for GJNs have been based on Reiman [37]. In [37] 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 [25] and studied in detail in [27]. In [9, 10] 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 [26], [38], 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 [7] 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 [5].

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 [29, 30] regarding the HT limits of the arrival flows and departure flows.

1.1.2. *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 [4], 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 [33]. Sigman [39, 40] 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 [8] for a different approach to stability via stochastic dominance. However, [40] and [8] 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 multi-class model, which was also studied in [42, 15]. (We do not consider the multi-class model here.) In [36] the stronger convergence in mean for queue length process and total workload process was established under slightly more restrictive conditions. In [28], 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.

1.2. *Organization.* The rest of the paper is organized as follows. We specify the model and establish the existence and convergence results for the stationary flows of a GJN in §2. We establish the main heavy-traffic limit for the stationary flows in §3.

We then establish more detailed results for three special cases in §4. First, we state the limit for the special case of a GJN with only one bottleneck queue, which is useful for the IDC approximations, because it involves only one-dimensional RBM. Corollary 4.3 shows that the approximation technique of feedback elimination is asymptotically correct in the HT limit. This extends the technique of immediate feedback elimination discussed in §III of [45].

In §5 we demonstrate how the HT limits can be used to derive approximations for the IDCs of the stationary flows, focusing on dependent superposition and slitting operations. These examples illustrate the complexity of the flows. The accuracy of the approximations in simulation comparisons also provide consistency checks for the HT limit theorems.

2. The Stationary Flows in an Open Queueing Network.

In this section, we establish the existence of the stationary flows in a GJN and convergence to those stationary flows. These issues can be complicated in general, but they are very manageable under appropriate regularity conditions, in particular, if we construct a Markov process representation and make assumptions implying Harris recurrence as in Chapter VII of [2] and [39, 40]. That allows the pre-limit process to be coupled with a stationary version, so that there is total variation convergence of the entire stochastic process. That implies convergence for a large class of related processes without complicated issues about the underlying topology.

In §2.1 we specify the model. Then in §2.2 we make assumptions implying the Harris recurrence and establish the existence 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 FCFS discipline. We associate with each station i an external arrival point process $A_{0,i}$ with finite rate

$$\lambda_{0,i} \equiv \lim_{t\to\infty} t^{-1} A_{0,i}(t) \tag{2.1}$$

where the limit holds w.p.1. Let $A_0 \equiv (A_{0,1}, \ldots, A_{0,K})$ denote the vector of all external arrival processes.

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

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

We assume that the service process $S_i(t)$ has finite rate μ_i, defined as in (2.1).

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 process, we let $\theta_i^l \in \{0,1\}^{K+1}$ indicate the routing vector of the l-th departure from queue i. So exactly one component of θ_i^l is 1 and the j-th component $\theta_{i,j}^l = 1$ indicates that the the l-th departure from the i-th station exits the system if $j = 0$ and is routed to station j if $1 \le j \le K$. Let

$$\Theta_i(n) \equiv (\Theta_{i,0}(n), \Theta_{i,1}(n), \ldots, \Theta_{i,K}(n)) = \sum_{l=1}^{n} \theta_i^l$$

denote the splitting decisions up to the n-th decision at station i. In the case of Markovian routing, the routing vectors are i.i.d., so that $\Theta_i(n)$ satisfies a functional weak law of large numbers (FWLLN), i.e.

$$\bar{\Theta}_{i,n}(t) \equiv \frac{1}{n}\Theta_i(\lfloor nt \rfloor) \Rightarrow \mathbf{p}_i t, \tag{2.2}$$

with the convergence uniform over bounded intervals.

Let $p_{i,j}$ denote the long run proportion of departures from station i that are routed to station j (assumed to exist). Let $P \equiv \{p_{i,j} : 1 \le i,j \le K\}$ be the routing matrix. Furthermore, let $p_{i,0} \equiv 1 - \sum_j p_{i,j}$ denote the fraction of customers that depart the system from station i.

To define the traffic intensities, we solve for the effective arrival rate at each node. 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 total arrival rate vector, which we obtain by solving the *traffic-rate equations*

$$\lambda_i = \lambda_{0,i} + \sum_{j=1}^{K} \lambda_{j,i} = \lambda_{0,i} + \sum_{i=1}^{K} \lambda_j p_{j,i}, \tag{2.3}$$

or in matrix form

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

where I denote the identity matrix. We assume that $I - P'$ is invertible; i.e., we assume that all customers eventually leave the system; see [11] or Theorem 3.2.1 of [32]. Hence, $\lambda_{i,j} \equiv \lambda_i p_{i,j}$ is the rate of the internal arrival stream from i to j.

For the internal arrival flows, let $A_{i,j}$ be the customer stream from i to j. Each internal arrival stream $A_{i,j}$ splits from the departure process D_i

according to the splitting decision process $\Theta_{i,j}$, so that

$$A_{i,j}(t) = \sum_{l=1}^{D_i(t)} \theta_{i,j}^l = \Theta_{i,j}(D_i(t)), \quad t \geq 0. \tag{2.4}$$

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 $A(t) \equiv (A_1(t), \ldots, A_K(t))$ collect all total arrival processes.

As observed in (7.1) and (7.2) in §7.2 of [9], the queue-length process is uniquely characterized by the flow balance equations for $1 \leq i \leq K$

$$Q_i(t) = Q_i(0) + A_i(t) - S_i(B_i(t)), \quad t \geq 0, \tag{2.5}$$

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

$$B_i(t) = \int_0^t 1_{Q_i(u)>0} du, \quad t \geq 0. \tag{2.6}$$

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 $D_{\text{ext}}(t) \equiv (D_{\text{ext},1}(t), \ldots, D_{\text{ext},K}(t))$ collect all external departure processes.

2.2. *Existence and Convergence Via Harris Recurrence.* In this section we establish the existence of the stationary flows and convergence to them 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:*

- *the external arrival processes are (possibly null) renewal processes with finite rates λ_i, interarrival times have finite squared coefficient of variation (scv) c^2_{a0i} for $1 \leq i \leq K$;*

- *the service times are i.i.d. with finite rates* μ_i *and finite scv* $c_{s_i}^2$ *for* $1 \leq i \leq K$*;*
- *the interarrival time or service time distributions have no mass at* 0*;*
- *the routing is Markovian with substochastic routing matrix* P*, so that* $I - P'$ *is invertible; and*
- *all processes are mutually independent.*

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 $\mathcal{S}(t) = (Q(t), U(t), V(t))$. Let $\mathcal{S}$ be an element of the function space $\mathcal{D}^{3K}$, i.e., with vectors of real-valued functions on the half-line $[0, \infty)$ that are right-continuous with left limits. Let the general initial condition be denoted by $\mathcal{S}(0) = (Q(0), U(0), V(0))$.

Given that we have a GJN, that means that the vector of external arrival processes A_0 will be a vector of delayed renewal process with the vector of first interarrival times being $U(0)$; the vector of service processes S will be a vector of delayed renewal process with first service time being $V(0)$; and the vector of queue length processes $Q(t)$ has a initial value of $Q(0)$. We refer to $\mathcal{S}$ as the *system state process.*

Now, define the auxiliary cumulative process $\mathcal{C}$, as in §VI.3 of [2], by

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

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

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

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

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

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

where

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

is a vector of cumulative point processes, with the processes defined in §2.1. We refer to $\mathcal{F}$ as the *flows.*

We say that the open queueing network is *stable* if there exist a stationary distribution π such that $\mathcal{S}(0) \sim \pi$ implies that $\mathcal{S}(t) \sim \pi$ for all $t \geq 0$. We say that a flow is *stationary* if it has stationary increments. We refer to [41] and Chapter 6 of [6] for background on stationary stochastic processes and ergodicity.

At this point we make the key assumption to obtain the Harris recurrence in [39, 40], [14] and Ch. VII of [2].

ASSUMPTION 2.2. *Each external interarrival-time distribution is unbounded above and spread out. That is, for external arrival process $A_{0,k}$ with interarrival distribution F_k, there exist a integer $j_k > 0$ such that the j_k-fold convolution $F_k^{*j_k}$ has an absolutely continuous component with respect to the Lebesgue measure.*

For a probability distribution to be spread out, it suffices for each interarrival-time distribution to have a positive probability density function (pdf) over some interval. That clearly avoids periodic behavior associated with the lattice case, but otherwise it is not restrictive for practical modeling. The unbounded condition could be replaced by the single external renewal arrival process with splitting in [40].

Finally, we assume that the queueing network is stable in the sense of the traffic intensities $\rho_i \equiv \lambda_i/\mu_i$, where λ_i is obtained from the traffic rate equations.

ASSUMPTION 2.3. *The traffic intensities satisfy* $\max_i \rho_i < 1$.

Under these three assumptions, Theorem 5.1 of [14] establishes stability of the GJN; also see Theorem 5.1 of [40] and [8, 23, 7] for alternative approaches and additional discussion.

THEOREM 2.1 (Harris recurrence from [14]). *Under Assumptions 2.1-2.3, the stochastic process $\mathcal{S}$ is a Harris recurrent Markov process.*

We now state the strong implications of Theorem 2.1. 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 $\mathcal{S}(s)$. Theorem 2.1 implies that

COROLLARY 2.1. *Under Assumptions 2.1-2.3, we have*

$$\mathcal{S}_s \Rightarrow \mathcal{S}_e \equiv (Q_e, U_e, V_e), \quad \textit{in} \quad \mathcal{D}^{3K} \quad \textit{as} \quad s \to \infty, \tag{2.11}$$

where $\mathcal{S}_e$ is a stationary process. Moreover, the convergence holds in the total variation metric, so that for any measurable function h from $\mathcal{D}^{3K}$ to a complete separable metric space,

$$h(\mathcal{S}_s) \Rightarrow h(\mathcal{S}_e) \quad \textit{as} \quad s \to \infty. \tag{2.12}$$

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, we define $A_{\mathrm{int},s}(t) = A_{\mathrm{int}}(t+s) - A_{\mathrm{int}}(s), A_s(t) = A(t+s) - A(s), D_s(t) = D(t+s) - D(s), D_{\mathrm{ext},s}(t) = D_{\mathrm{ext}}(t+s) - D_{\mathrm{ext}}(s)$, $B_s(t) = B(t+s) - B(s)$ 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 as

$$S_s(t) \equiv S(B(s)+t) - S(B(s)), \tag{2.13}$$

which is a vector of delayed renewal processes with first intervals distributed as $V(s)$, the residual service time at time s. This definition of the service process allow us to write the departure process as a composition of the two processes S_s and B_s via

$$\begin{aligned} D_s(t) \equiv D(s+t) - D(s) &= S(B(s+t)) - S(B(s)) \\ &= S_s(B_s(t)) \equiv (S_s \circ B_s)(t), \quad t \geq 0. \end{aligned} \tag{2.14}$$

Finally, let $\mathcal{C}_s \equiv (B_s, Y_s)$ and $\mathcal{F}_s \equiv (A_{0,s}, A_{\mathrm{int},s}, A_s, S_s, D_s, D_{\mathrm{ext},s})$.

Theorem 2.2 (Existence and convergence of the stationary flows). *Under Assumptions 2.1-2.3, there exists a unique stationary and ergodic cumulative processes (with stationary increments satisfying the LLN)*

$$\mathcal{C}_e \equiv (B_e, Y_e)$$

and

$$\mathcal{F}_e \equiv (A_{0,e}, A_{\mathrm{int},e}, A_e, S_e, D_e, D_{\mathrm{ext},e})$$

and a unique stationary process

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

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

$$\mathcal{M}_s \equiv (\mathcal{S}_s, \mathcal{C}_s, \mathcal{F}_s) \Rightarrow (\mathcal{S}_e, \mathcal{C}_e, \mathcal{F}_e) \equiv \mathcal{M}_e \quad \text{in} \quad \mathcal{D}^{10K+K^2}. \tag{2.15}$$

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 $V_e(0)$ and the mode of convergence can be strengthened to in total variation.

Proof. By Corollary 2.1 and the definition of S_s in (2.13), the convergence of $V_s(0) = V(s)$ implies the convergence of S_s to S_e, which is a delayed renewal process with first interval distributed as $V_e(0)$ and other intervals distributed as a generic service time. By Assumption 2.1, $A_{0,s}$ converges to $A_{0,e}$. Hence, we have as $s \to \infty$

$$(Q_s, U_s, V_s, A_{0,s}, S_s) \Rightarrow (Q_e, U_e, V_e, A_{0,e}, S_e) \quad \text{in} \quad \mathcal{D}^{5K}, \tag{2.16}$$

where the subscript e denote the stationary versions.

For the cumulative busy time process defined in (2.6), note that

$$B_{i,e}(t) = \int_0^t 1_{Q_{i,e}(u)>0} du, \tag{2.17}$$

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 [6].) Moreover, without having to carefully consider continuity, we have

$$B_{i,s}(t) = \int_s^{s+t} 1_{Q_i(u)>0} du = \int_0^t 1_{Q_{i,s}(u)>0} du.$$

Hence, we can extend the convergence as $s \to \infty$ in (2.16) to

$$(Q_s, U_s, V_s, A_{0,s}, S_s, B_s, Y_s) \Rightarrow (Q_e, U_e, V_e, A_{0,e}, S_e, B_e, Y_e) \tag{2.18}$$

in $\mathcal{D}^{7K}$. For the departure process, recall from (2.14) that $D_s(t) = S_s(B_s(t))$, so that we can apply the the composition map and (2.18) to obtain as $s \to \infty$

$$(Q_s, U_s, V_s, A_{0,s}, S_s, B_s, Y_s, D_s) \Rightarrow (Q_e, U_e, V_e, A_{0,e}, S_e, B_s, Y_e, D_e) \tag{2.19}$$

in $\mathcal{D}^{8K}$. Similarly, jointly with the limits above, we can add limit for other processes. For the total arrival process, we have

$$\begin{aligned} A_{i,s}(t) = A_i(s+t) - A_i(s) &= D_{i,s}(t) + Q_i(s+t) - Q_i(s) \\ &\Rightarrow D_{i,e}(t) + Q_{i,e}(t) - Q_{i,e}(0) \quad \text{as} \quad s \to \infty. \end{aligned}$$

So we have convergence if we define $A_e \equiv D_{i,e}(t) + Q_{i,e}(t) - Q_{i,e}(0)$.

For internal arrival process, by definition (2.4),

$$A_{i,j,s}(t) = A_{i,j}(t+s) - A_{i,j}(s) = \Theta_{i,j}(D_i(t+s)) - \Theta_{i,j}(D_i(s)).$$

Under Markovian routing, the right-hand-side above is in distribution equivalent to $\Theta_{i,j}(D_i(t+s) - D_i(s)) = \Theta_{i,j}(D_{i,s}(t))$. Hence, as $s \to \infty$,

$$A_{s,\text{int}} \Rightarrow A_{e,\text{int}} \equiv (\Theta_{i,j}(D_{i,e}(t)) : 1 \le i, j \le K).$$

Similarly, we can add external departure processes to the limit, with

$$D_{\text{ext},e} \equiv (\Theta_{i,0}(D_{i,e}(t)) : 1 \le i \le K). \quad \blacksquare$$

3. Heavy-Traffic Limit Theorems for the Stationary Processes. To set the stage for our heavy-traffic limits, in §3.1 we present a centered representation of the flows. This representation parallels those used in [37, 9, 10, 14], 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 $\lambda = (I - P')\lambda_0$ as in (2.3). For service, we start with rate-1 base service process S_i^0 for station i and scale it by μ_i so that the service process at station i is denoted by $S_i \equiv S_i^0 \circ \mu_i e$ with $e(t) = t$ being the identity function. Let the center processes be defined by

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

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

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

where Y is defined in (2.8).

The next theorem expresses the queue length processes, the centered total arrival and the centered departure flows in terms of the centered external arrival, service and routing processes. Let ψ be the K-dimensional reflection map; e.g., see Chapter 14 of [47].

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

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

while the queue length process can be written as

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

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

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

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

where the centered processes are defined in (3.1).

Remark 3.1 (Stationary flows). *Note that the representation in Theorem 3.1 does not impose any assumption on the initial condition of the open queueing network. As ensured by Theorem 2.2, there exists a stationary distribution π such that the flows are stationary if $\mathcal{S}(0) \sim \pi$. With this specific initial condition, Theorem 3.1 applies to the stationary flows.*

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

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

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 $Q_i = 0$, (3.4) follows from the usual reflection argument.

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

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

or, in matrix notation, by

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

By (3.4), we have

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

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

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

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

3.2. *Heavy-Traffic Limit with Any Subset of Bottlenecks.* Throughout this section, we assume that the system is stationary in the sense of Theorem 2.2 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 heavy-traffic limit while other stations stay unsaturated. Two important special cases are: (i) $|\mathcal{H}| = K$ so that all stations approaches heavy traffic at the same time, which corresponds to the original case in [37]; and (ii) $|\mathcal{H}| = 1$ so that only one station is in heavy traffic. 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 $\mu_{i,\rho} = \lambda_i/(c_i\rho)$ for $1 \le i \le K$ and set $c_i = 1$ for all $i \in \mathcal{H}$ and $c_i < 1$ for all $i \notin \mathcal{H}$. 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's in (3.3) are now specified as above.

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

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

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 flows, defined as

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

Finally, let $W^*_{i,\rho}(t) \equiv (1-\rho)W_{i,\rho,\lfloor(1-\rho)^2 t\rfloor}$ denote the HT scaled waiting time process, where $W_{i,\rho,n}$ denotes the waiting time of the n-th customer at station i in the ρ-th system; and 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 notations by discussing a modified and yet asymptotically equivalent system, where all service times at the nonbottleneck queues are set to zero.

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

$\hat{P}_{\mathcal{H}}$, is

$$\begin{aligned} \hat{P}_{\mathcal{H}} &= P_{\mathcal{H}} + \sum_{l=0}^{\infty} P_{\mathcal{H},\mathcal{H}^c} \left(P_{\mathcal{H}^c}\right)^l P_{\mathcal{H}^c,\mathcal{H}} \\ &= P_{\mathcal{H}} + P_{\mathcal{H},\mathcal{H}^c} \sum_{l=0}^{\infty} \left(P_{\mathcal{H}^c}\right)^l P_{\mathcal{H}^c,\mathcal{H}} \\ &= P_{\mathcal{H}} + P_{\mathcal{H},\mathcal{H}^c} \left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c,\mathcal{H}}. \qquad (3.9) \end{aligned}$$

Note that the inverse $\left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1}$ appearing in (3.9) is the fundamental matrix associated with the transient finite Markov chain with transition matrix $P_{\mathcal{H}^c}$. If we let $\hat{P}_{\mathcal{H}^c,\mathcal{H}}$ denote the matrix of the probabilities that the first visit to a bottleneck queue of an external arrival at a non-bottleneck queue $i \in \mathcal{H}^c$ is at $j \in \mathcal{H}$, then we have

$$\hat{P}_{\mathcal{H}^c,\mathcal{H}} = \sum_{l=0}^{\infty} (P_{\mathcal{H}^c})^l P_{\mathcal{H}^c,\mathcal{H}} = \left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c,\mathcal{H}}. \qquad (3.10)$$

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

$$\hat{\lambda}_{0,\mathcal{H}} = \lambda_{0,\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} \lambda_{0,\mathcal{H}^c} = \lambda_{0,\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}} \left(I_{\mathcal{H}^c} - P'_{\mathcal{H}^c}\right)^{-1} \lambda_{0,\mathcal{H}^c}, \qquad (3.11)$$

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

$$\hat{\lambda}_{\mathcal{H}} = (I - \hat{P}'_{\mathcal{H}})^{-1} \hat{\lambda}_{0,\mathcal{H}} = \lambda_{\mathcal{H}}. \qquad (3.12)$$

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

The following theorem states the joint heavy-traffic limit of the queue length process, the workload and waiting time processes, the splitting-decision process and all the flows. As in [9, 10], we allow an arbitrary subset of nodes to be bottleneck queues (critically loaded) while the rest are sub-critically loaded. To treat the stationary processes, we apply [23] and [7], extended to include non-bottleneck queues. Because our basic model data involves only single arrival and service processes, with only the parameters being scaled, we do not need Assumption (A4) in [7].

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 $\rho \uparrow 1$,

$$\begin{aligned}&(Q^*_\rho, W^*_\rho, Z^*_\rho, \Theta^*_\rho, \Theta^*_{\mathrm{ext},\rho}, \mathcal{F}^*_\rho)\\ &\quad \Rightarrow (Q^*, W^*, Z^*, \Theta^*, \Theta^*_{\mathrm{ext}}, \mathcal{F}^*) \quad \textit{in} \quad \mathcal{D}^{9K+2K^2},\end{aligned} \tag{3.13}$$

where:

1. *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. $(\Theta^*_{i,j} : 0 \le j \le K)$ 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 $(\Theta^*_{i,j} : 0 \le j \le K)$ are mutually independent, $1 \le i \le K$.*
2. *The queue length process Q^* consists of two parts. $Q^*_{\mathcal{H}^c} \equiv 0$ and $Q^*_{\mathcal{H}}$ is a stationary $|\mathcal{H}|$-dimensional RBM*

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

where $\psi_{\mathcal{H}}$ is the $|\mathcal{H}|$-dimensional refelction map with reflection matrix $R_{\mathcal{H}} \equiv I - \hat{P}_{\mathcal{H}}$ and $\hat{X}^*_{\mathcal{H}}$ is the net-input process associated with the bottleneck queues, defined below. Furthermore, $Q^*_{\mathcal{H}}(0)$ has unique stationary distribution of the stationary RBM. $\hat{X}^*_{\mathcal{H}}$ is a $|\mathcal{H}|$-dimensional Brownian motion

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

where $e_{\mathcal{A}}$ collects columns in the K-dimensional identity matrix I that corresponds to index set $\mathcal{A}$; $\hat{P}_{\mathcal{H}}$, $\hat{P}_{\mathcal{H}^c,\mathcal{H}}$ and $\hat{\lambda}_{0,\mathcal{H}}$ are defined in (3.9), (3.10) and (3.11), respectively.
3. *The total arrival process A^* can be regarded as a stationary process, having stationary increments, specified by*

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

4. *The stationary departure process D^* is specified as*

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

In particular,

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

5. *The internal arrival flow $A^*_{i,j}$ can be expressed as*

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

and the external departure flow can be expressed as

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

6. $Z^*_i = \lambda_i^{-1} Q^*_i$ *and* $W^*_i = Z^*_i \circ \lambda_i e$.

Proof of Theorem 3.2. Much of the statement follows from [9, 10] and [7]. First, the HT limit for the state process with an arbitrary subset $\mathcal{H}$ of critically loaded stations follows from [9, 10]. Second, the HT limit for the steady-state queue length follows from [7]. The papers [23] and [7] do not consider non-bottleneck stations, but their arguments extend to that more general setting. (See Remark 3.3 below for discussion.) We subsequently establish the heavy-traffic limits for the flows. We do so by exploiting the continuous mapping theorem with the direct representations of the stationary flows that we have established.

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

$$X^*_\rho = Q^*_\rho(0) + A^*_{0,\rho} + \left(\tilde{\Theta}^*_\rho\right)' \mathbf{1} - (I - P')\tilde{S}^*_\rho + (\lambda_0 - (I - P')\mu_\rho)(1-\rho)^{-1} e, \tag{3.14}$$

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

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

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

Solving for $Y^*_{\mathcal{H}^c,\rho}$ in (3.16) and substituting into (3.15), we have

$$Q^*_{\mathcal{H},\rho} = \hat{X}^*_{\mathcal{H},\rho} + (I - \hat{P}'_{\mathcal{H}})Y^*_{\mathcal{H},\rho}, \tag{3.17}$$

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

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

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

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

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

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

$$\eta_{\mathcal{H}^c,\rho} = \lambda_{0,\mathcal{H}^c} - (I - P'_{\mathcal{H}^c,\mathcal{H}^c})\mu_{\mathcal{H}^c,\rho} + P'_{\mathcal{H},\mathcal{H}^c}\mu_{\mathcal{H},\rho}. \tag{3.19}$$

Hence

$$\begin{aligned}
\hat{\eta}_\rho & \equiv \eta_{\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\eta_{\mathcal{H}^c,\rho} \\
& = \lambda_{0,\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\lambda_{0,\mathcal{H}^c} - (I - \hat{P}'_{\mathcal{H}})\mu_{\mathcal{H},\rho}.
\end{aligned} \tag{3.20}$$

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 (3.20), we have

$$\hat{\eta}_\rho = (I - \hat{P}'_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho}). \tag{3.21}$$

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

$$\begin{aligned}\hat{X}^*_{\mathcal{H},\rho} \;=\; & Q^*_{\mathcal{H},\rho}(0) + A^*_{0,\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1} A^*_{0,\mathcal{H}^c,\rho} - (I - \hat{P}'_{\mathcal{H}})\tilde{S}^*_{\mathcal{H},\rho} \\ & + e'_{\mathcal{H}}(\tilde{\Theta}^*_\rho)'\mathbf{1} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1} e'_{\mathcal{H}^c}(\tilde{\Theta}^*_\rho)'\mathbf{1} \\ & + (I - \hat{P}_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho})(1-\rho)^{-1} e \\ & + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q^*_{\mathcal{H}^c,\rho}(0) - Q^*_{\mathcal{H}^c,\rho}).\end{aligned} \tag{3.22}$$

Now we are ready to deduce the claimed conclusions. First for conclusion 1, most follows directly from Donsker's theorem, Theorem 4.3.2 of [47], 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 [10].

For conclusion 2, we apply [7] to get

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

Then we can apply the representation (3.22) we have just derived above plus the continuous mapping theorem to obtain the conclusion, as in [10].

Then the conclusion 2 follows from Theorem 6.1 of [10]. 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 $\rho \to 1$,

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

and

$$Q^*_{\mathcal{H},\rho} = \hat{X}^*_{\mathcal{H},\rho} + (I - \hat{P}'_{\mathcal{H}})Y^*_{\mathcal{H},\rho} = \psi_{I-\hat{P}'_{\mathcal{H}}}(\hat{X}^*_{\mathcal{H},\rho}). \tag{3.23}$$

Conclusions 3 and 4 follows from the representations derived in Theorem 3.1, the continuous mapping theorem and the established convergence of the queue length process, the external arrival processes and the splitting-decision processes. To this end, we only need to apply diffusion scaling (accelerate time by $(1-\rho)^{-2}$ and scale space by $(1-\rho)$) to the representations in Theorem 3.1 so that

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

$$D_\rho^* = (I - P')^{-1} \left(Q_\rho^*(0) - Q_\rho^* + A_{0,\rho}^* + (\tilde{\Theta}_\rho^*)'\mathbf{1} \right). \tag{3.24}$$

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

Next, conclusions 5 follows from the limit of the departure process and the FCLT of the splitting operation in §9.5 of [47]. Finally, the associated limits for the waiting time and workload can be related to the limit for the queue length as indicated in [10]. ■

Remark 3.3 (Elaboration on the application of [7]). We apply [7], 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, 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 (3.17) and (3.22) can be mapped to the representations (16) on p.51 of [7], but with slightly more complicated constant terms associated with the matrix multiplication we have in (3.22). Noting the expression of the drift term we have in (3.21), the rest of the proof is essentially the same. For $Q_{\mathcal{H}^c}^*$, by [9, 10], it is negligible in the sense of Theorem 3.3 of [7]. Theorem 3.4 of [7] relies only on the moment estimation as in their Theorem 3.3 and the Markov property of $\mathcal{S}(t)$ (which they denoted as $X(t)$). Finally, Theorem 3.5 and Theorem 3.2 of [7] remain unchanged.

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, see 5 for more details.

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. *Single-Server Queue with i.i.d. 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.

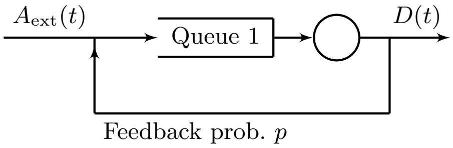

FIG 1. *A single-server queue with feedback example.*

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 $\rho \uparrow 1$,*

$$(Q^*_\rho, W^*_\rho, Z^*_\rho, \Theta^*_\rho, \Theta^*_{\text{ext},\rho}, \mathcal{F}^*_\rho) \Rightarrow (Q^*, W^*, Z^*, \Theta^*, \Theta^*_{\text{ext}}, \mathcal{F}^*) \quad \text{in} \quad \mathcal{D}^{11},$$

where $\mathcal{F}^*_\rho = (A^*_{0,\rho}, A^*_\rho, A^*_{\text{int},\rho}, S^*_\rho, D^*_\rho, D^*_{\text{ext},\rho})$, $\mathcal{F}^* = (A^*_0, A^*, A^*_{\text{int}}, S^*, D^*, D^*_{\text{ext}})$ and:

1. *$A^*_0 = c_{a_0} B_{a_0} \circ \lambda_0 e$ and $S^* = c_s B_s \circ \lambda e$, where B_{a_0} and B_s are standard Brownian motions. $(\Theta^*, \Theta^*_{\text{ext}})$ is a zero-drift two-dimensional Brownian motion with covariance matrix $\Sigma = (\sigma^2_{i,j} : 1 \le i, j \le 2)$, where $\sigma^2_{1,1} = \sigma^2_{2,2} = p(1-p)\lambda$ and $\sigma^2_{1,2} = \sigma^2_{2,1} = -p(1-p)\lambda$, so that*

$$\Theta^* + \Theta^*_{\text{ext}} = 0.$$

   Furthermore, B_{a_0}, B_s and $(\Theta^*, \Theta^*_{\text{ext}})$ are mutually independent.
2. *The queue length process Q^* is a stationary one-dimensional RBM*

$$Q^* \equiv \psi\left(X^*\right),$$

   where ψ is the one-dimensional reflection map and X^* is a one-dimensional Brownian motion

$$X^* = Q^*(0) + A^*_0 + (\Theta^* - (1-p)S^*) - \lambda_0 e.$$

   Furthermore, $Q^*(0)$ 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 $c_x^2/2$.

3. *The total arrival process A^* can be regarded as a stationary process, having stationary increments, specified by*

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

4. *The stationary total departure process D^* is specified as*

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

5. *The internal arrival flow A^*_{int} can be expressed as*

$$A^*_{\text{int}} = pD^* + \Theta^*$$

and the external departure flow can be expressed as

$$D^*_{\text{ext}} = (1-p)D^* + \Theta^*_{\text{ext}} = A_0^* + Q^*(0) - Q^*.$$

6. *$Z^* = \lambda^{-1}Q^*$ and $W^* = Z^* \circ \lambda e$.*

Remark 4.1 (Eliminating immediate feedback). As observed in Section III of [45], 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.

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. For waiting time, we need to adjust for per-visit waiting time by multiplying the waiting time in the modified system by $(1-p)$. This modification results in a change in service scv, by conditional variance formula, the scv of the total service time is $\tilde{c}_s^2 = p + (1-p)c_s^2$. From the aspect of the FCLT, let $\tilde{S}^* \equiv \Theta^* - (1-p)S^*$, we argue that $\tilde{S}^*$ has the same distribution as the diffusion limit of the new service counting process. To this end, note that $\Theta^* = \sqrt{p(1-p)}B_\Theta \circ \lambda e$ and $S^* = c_s B_s \circ \lambda e$, where B_Θ and B_s are independent standard Brownian motions (zero drift and unit variance). Hence, from part (ii) of Corollary 4.1, the stationary net-input process of the original system is

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

where $\tilde{S}^* \overset{dist.}{=} \tilde{c}_s \tilde{B}_s \circ \lambda_0 e$, $\tilde{B}_s$ is a standard Brownian motion and $\lambda_0 = (1-p)\lambda$. On the other hand, (4.1) is exactly the diffusion limit of the net-input process of a single-server queue with arrival process A_0 and service process $\tilde{S}$. Starting from the equivalence of the net-input process, we obtain the equivalence of the queue length process by part (ii) of Corollary 4.1, which in turn gives the equivalence of the external departure process as well as the waiting time and workload process.

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 2.1-2.3, consider a series of GJNs in stationarity, indexed by ρ. Assume that $\mu_{i,\rho} = \lambda_i/(c_i\rho)$ for $1 \leq i \leq K$ and set $c_h = 1$ and $c_i < 1$ for all $i \neq h$. Then, we have*

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

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

1. *For $0 \leq i \leq K$, $A^*_{0,i} = c_{a_{0,i}} B_{a_{0,i}} \circ \lambda_{0,i} e$ 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. $(\Theta^*_{i,j} : 0 \leq j \leq K)$ is a zero-drift $(K+1)$-dimensional Brownian motion with covariance matrix $\Sigma_i = (\sigma^2_{j,k} : 0 \leq j,k \leq K)$, 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 $0 \leq i \neq j \leq K$. Furthermore, $B_{a_{0,i}}$, B_{s_i} and $(\Theta^*_{i,j} : 0 \leq j \leq K)$ are mutually independent, $1 \leq i \leq K$.*
2. *The queue length process Q^* consists of two parts. $Q^*_i \equiv 0$ for $i \neq h$ and Q^*_h is a stationary one-dimensional RBM*

$$Q^*_h \equiv \psi\left(\hat{X}^*_h\right),$$

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

$$\hat{X}^*_h = Q^*_h(0) + A^*_{0,h} + \hat{P}'_{\mathcal{H}^c,h} A^*_{0,\mathcal{H}^c} + e'_h (\Theta^*)' \mathbf{1} + \hat{P}'_{\mathcal{H}^c,h} e'_{\mathcal{H}^c} (\Theta^*)' \mathbf{1}. \tag{4.2}$$

$$- (1-\hat{P}_h) S^*_h - \hat{\lambda}_{0,h} e \tag{4.3}$$

where $e_{\mathcal{A}}$ collects columns in the K-dimensional identity matrix I that corresponds to index set $\mathcal{A}$; $\hat{P}_{\mathcal{H}}$, $\hat{P}_{\mathcal{H}^c,\mathcal{H}}$ and $\hat{\lambda}_{0,\mathcal{H}}$ are defined in (3.9), *(3.10) and (3.11), respectively. Furthermore, $Q_h^*(0)$ has unique stationary distribution of the stationary RBM.*

3. *The total arrival process A^* can be regarded as a stationary process, having stationary increments, specified by*

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

4. *The stationary departure process D^* is specified as*

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

In particular,

$$D_{\mathcal{H}^c}^* = Q_{\mathcal{H}^c}^* + A_{\mathcal{H}^c}^* - Q_{\mathcal{H}^c}^*(0) = A_{\mathcal{H}^c}^*.$$

5. *The internal arrival flow $A_{i,j}^*$ can be expressed as*

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

and the external departure flow can be expressed as

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

6. *$Z_i^* = \lambda_i^{-1} Q_i^*$ and $W_i^* = Z_i^* \circ \lambda_i e$.*

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 *(3.10)* and recall that $\hat{p} \equiv \hat{P}_h$ is the feedback probability defined in Remark 3.2.

THEOREM 4.1. *The HT limit $\hat{X}_h^*$ in (4.3) can be expressed as the following one-dimensional Brownian motion*

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

where

$$\hat{A}^* = A_{0,h}^* + \sum_{i \in \mathcal{H}^c} \left(\hat{p}_{i,h} A_{0,i}^* + \hat{\Theta}_{i,h}^* \right), \tag{4.4}$$

and

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

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

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

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

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

Now, consider a reduced one-station network consist of the only bottleneck queue, while all non-bottleneck queues have service times set to 0 so that they serve as instantaneous switches. In the reduced network, we define an external arrival $\hat{A}_0$ to the bottleneck queue to be any external arrival that arrive at the bottleneck queue for the first time. Hence, an external arrival may have visited one or multiple non-bottleneck queues before its first visit to the bottleneck queue. In particular, the external arrival process can be expressed as the superposition of (i) the original external arrival process $A_{0,h}$ at station h; and (ii) the Markov splitting of the external arrival process $A_{0,i}$ at station i with probability $\hat{p}_{i,h}$, for $i \in \mathcal{H}^c$.

Theorem 4.1 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 4.1 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 Remark 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 time is exactly the geometric sum of the original service times at the bottleneck queue. Theorem 4.1 also implies that the service process

$$\hat{S}^* \equiv \hat{\Theta}^*_S - (1-\hat{p})S^*_h, \tag{4.5}$$

shares the same diffusion limit with a modified service process after feedback elimination.

Hence, we have the following corollary.

Corollary 4.3 (Feedback elimination with one bottleneck queue). *Eliminating all feedback at the bottleneck queue as described above prior to analysis is asymptotically correct in HT for GJNs with a single bottleneck queue.*

4.3. *Functional Central Limit Theorem for the Stationary Flows.* In this section, we focus on yet another important specail case of Theorem 3.2 where we set $|\mathcal{H}| = 0$. In this special case, all stations are strictly non-bottleneck, i.e., $\mu_{i,\rho} = \lambda/(c_i\rho)$ where $c_i < 1$ for all i. As $\rho \uparrow 1$, the family of systems converges to a limiting system where the traffic intensity at station i is $\rho_i = c_i$. Hence, the scaling used in (3.8) corresponds to the diffusion scaling used in the usual FCLT. The following corollary describes the joint FCLT of the stationary flows.

Corollary 4.4 (FCLT for the stationary flows). *Under Assumption 2.1-2.3, consider a family of open queueing networks in stationarity, indexed by ρ. Assume that $\mu_{i,\rho} = \lambda_i/(c_i\rho)$ with $c_i < 1$ for $1 \le i \le K$. Then, as $\rho \uparrow 1$,*

$$\begin{aligned} &(Q^*_\rho, W^*_\rho, Z^*_\rho, \Theta^*_\rho, \Theta^*_{\text{ext},\rho}, \mathcal{F}^*_\rho) \\ &\quad \Rightarrow (Q^*, W^*, Z^*, \Theta^*, \Theta^*_{\text{ext}}, \mathcal{F}^*) \quad \text{in} \quad \mathcal{D}^{9K+2K^2}, \end{aligned} \tag{4.6}$$

where:

1. *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. $(\Theta^*_{i,j} : 0 \le j \le K)$ 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 $(\Theta^*_{i,j} : 0 \le j \le K)$ are mutually independent, $1 \le i \le K$.*
2. *The queue length process $Q^* \equiv 0$.*

3. *The total arrival process A^* can be regarded as a stationary process, having stationary increments, specified by*

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

4. *The stationary departure process is the same as the stationary total arrival process, so that $D^* = A^*$.*
5. *The internal arrival flow $A_{i,j}^*$ can be expressed as*

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

and the external departure flow can be expressed as

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

6. *Finally, $Z_i^* = W_i^* = 0$.*

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.

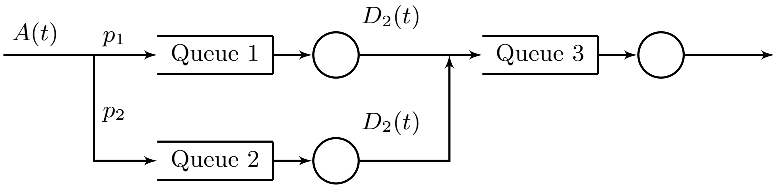

FIG 2. *A re-combining after splitting example.*

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*

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

where

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

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

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

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

$$\begin{aligned} \mathrm{cov}(D_{1,\rho}(t), D_{1,\rho}(t)) &= \mathrm{cov}\left((1-\rho_1)^{-1}D_{1,\rho}^*((1-\rho_1)^2t), (1-\rho_1)^{-1}D_{2,\rho}^*((1-\rho_1)^2t)\right) \\ &\Rightarrow (1-\rho_1)^{-2}\mathrm{cov}(D_1^*((1-\rho_1)^2t), D_2^*((1-\rho_1)^2t)), \end{aligned}$$

as $\rho \uparrow 1$.

On the other hand, by applying Corollary 5.1 of [51], we have

$$\begin{aligned} \mathrm{cov}\left(D_1^*(t), D_2^*(t)\right) &= \mathrm{cov}(A_1^*(t), A_2^*(t)) - \mathrm{cov}(Q_1^*(t), A_2^*(t)) \\ &= p_1(1-p_1)(c_{a_0}^2-1)\lambda t - \mathrm{cov}(Q_1^*(t)), A_2^*(t)) \\ &= p_1(1-p_1)(c_{a_0}^2-1)\lambda t w^*\left(p_1\lambda t/c_{x_1}^2\right), \end{aligned}$$

where $c_{x_1}^2 = c_{a_1}^2 + c_s^2$, $c_{a_1}^2 = p_1c_a^2 + (1-p_1)$ and w^* is the weight function defined in (28) of [51]. 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 $c_a^2 = 4$, $p_1 = 0.25$, $p_2 = 0.75$ and i.i.d. Erlang (E_2) service times with $c_{s_i}^2 = 0.5$. 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.

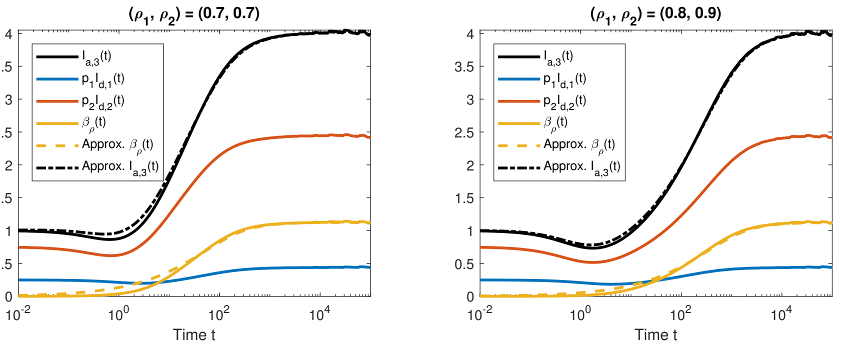

FIG 3. *Two examples in Example 5.1.*

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

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

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

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

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

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

so that

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

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 $c_{x_1}^2 = c_a^2 + c_s^2$, $c_a^2 = \frac{1}{1-p}c_{a0}^2 + \frac{p}{1-p}$, w^* is the weight function defined in (28) of [51] and the explicit expression is derived using Corollary 5.1 of [51].

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

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

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

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

with

$$\begin{aligned}
\beta_\rho(t) &\equiv 2\text{cov}(A_0^*((1-\rho)^2 t), A_{\text{int}}^*((1-\rho)^2 t))/(\lambda(1-\rho)^2 t) \\
&= 2pc_{a_0}^2 w^*((1-\rho)^2/c_x^2),
\end{aligned} \tag{5.8}$$

where again $c_x^2 = c_a^2 + c_s^2$ and $c_a^2 = \frac{1}{1-p}c_{a0}^2 + \frac{p}{1-p}$.

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 $c_s^2 = 4$. The plot on the left focuses on the feedback flow $A_{\text{int}}(t)$, while the plot on the right focuses on the superposition arrival process $A(t)$. Again, the approximation matches simulation remarkably well.

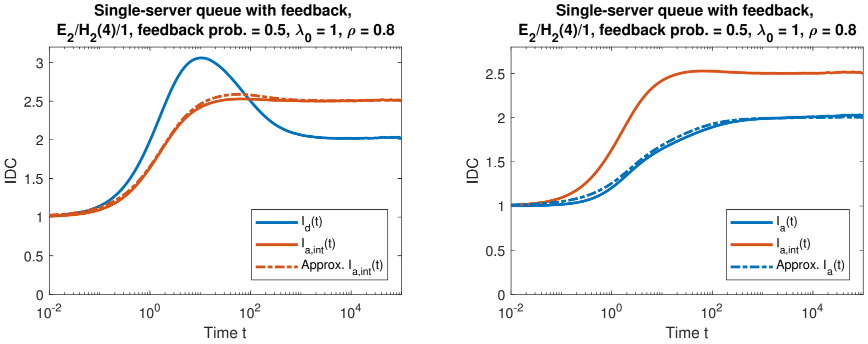

Fig 4. *Left plot shows the dependent splitting in a single-server queue with feedback example. Model parameters are described in the title. The simulation estimation of the IDC of the feedback flow is contrasted to the IDC approximation* (5.5) *with correction term* (5.6) *in dotted-and-dashed lines. 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.*

Acknowledgements. We thank Karl Sigman for helpful discussion about Harris recurrence. We received support from NSF grant CMMI 1634133.

REFERENCES

[1] J. Abate and W. Whitt. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems*, 10:5–88, 1992.

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

[3] F. Baccelli and P. Bremaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences.* Springer, New York, second edition, 2003.

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

[5] 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.

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

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

[8] 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.

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

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

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

[12] D. R. Cox. *Renewal Theory.* Methuen, London, 1962.

[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] J. G. Dai and J. M. Harrison. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *The Annals of Applied Probability*, pages 65–86, 1992.

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

[19] 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.

[20] 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.

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

[22] K. W. Fendick and W. Whitt. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE*, 71(1):171–194, 1989.

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

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

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

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

[28] 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.

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

[30] 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.

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

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

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

[34] P. J. Kuehn. Approximate analysis of general queuing networks by decomposition. *IEEE Transactions on Communications*, 27(1):113–126, 1979.

[35] 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.

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

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

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

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

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

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

[42] 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.

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

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

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

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

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

[48] W. Whitt and W. You. The advantage of indices of dispersion in queueing approximations. Under review, Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2018.

[49] W. Whitt and W. You. A robust queueing network analyzer based on indices of dispersion. Under review, Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 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] Ward Whitt and Wei You. Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function. *Stochastic Systems*, 8(2):143–165, 2018.

Department of IEOR,
S. W. Mudd Building,
500 West 120th Street,
New York, NY 10027-6699
E-mail: ww2040@columbia.edu
wy2225@columbia.edu