Queueing Systems (2023) 103:313–346
https://doi.org/10.1007/s11134-022-09869-1

Diffusion parameters of flows in stable multi-class queueing networks

Sarat Babu Moka[1] · **Yoni Nazarathy**[2] · **Werner Scheinhardt**[3]

Received: 25 August 2020 / Revised: 5 November 2022 / Accepted: 9 November 2022 /
Published online: 24 November 2022

Abstract
We consider open multi-class queueing networks with general arrival processes, general processing time sequences and Bernoulli routing. The network is assumed to be operating under an arbitrary work-conserving scheduling policy that makes the system stable. We study the variability of flows within the network. Computable expressions for quantifying flow variability have previously been discussed in the literature. However, in this paper, we shed more light on such analysis to justify the use of these expressions in the asymptotic analysis of network flows. Toward that end, we find a simple diffusion limit for the inter-class flows and establish the relation to asymptotic (co-)variance rates.

Keywords Queueing networks · Diffusion limits · Asymptotic variance

Mathematics Subject Classification 60K25

1 Introduction

The study of explicit performance measures of stable queueing networks has been at the heart of applied probability and operations research for the past half century. Initial results such as Burke's theorem [7], indicating that the output of a stationary M/M/1 queue is a Poisson process, have motivated the study of queueing output processes with the aim of using the output characteristics of one queue as the input of a downstream queue. While landmark results such as the product form solution of Jackson networks (cf. [23, 26]) have given much hope and practical utility, in the 1960s and 1970s it was

✉ Sarat Babu Moka
sarat.moka@mq.edu.au

1 School of Mathematical and Physical Sciences, Macquarie University, Sydney, Australia

2 School of Mathematics and Physics, The University of Queensland, Brisbane, Australia

3 Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

well understood that explicit exact queueing network decomposition is in general not attainable.

The lack of explicit solutions in general cases as well as the inability to exactly decouple most networks has motivated the study of queueing output processes as in [3, 14–16, 22, 38]. That line of work is also coupled with the development and study of heuristic queueing network decomposition schemes such as the queueing network analyzer (QNA) [36] (see also [30]) and many subsequent approximation methods (see for example the heuristics in [28]). The typical approximating assumption made in such schemes is that each queue in isolation is a G/G/1 queue which can be analyzed independently of the other queues. The input process is then approximated by taking into consideration both exogenous arrivals and departures from other queues in the network.

Some of the key ingredients needed for a network decomposition (such as QNA) are based on

$$\lambda_k := \lim_{t\to\infty} \frac{\mathbb{E}[A_k(t)]}{t}, \quad \text{and} \quad \sigma_k^2 = \lim_{t\to\infty} \frac{\text{Var}\Big(A_k(t)\Big)}{t},$$

where $A_k(t)$ is the arrival counting process into the queue of class k:

$$A_k(t) := E_k(t) + \sum_i D_{i,k}(t),$$

with $E_k(t)$ representing the exogenous arrival counting process to that queue and $D_{i,j}(t)$ the number of items that have departed from queue i and immediately arrived to queue j during the time interval $[0, t]$. We refer to the counting process $D_{i,j}(t)$ as *flow* $i \to j$. The summation in $A_k(t)$ is over all flows $i \to k$.

Finding λ_k exactly is typically a trivial matter based on the network routing matrix and exogenous arrival rates. As opposed to that, σ_k^2 is more complex. In fact, computable expressions for σ_k^2 have only been presented as part of the so-called innovations method in [28]. Here, the author builds on earlier work [29] by Kim, Muralidharan and O'Cinneide and presents an expression that yields σ_k^2 among other performance measures (see equation (42) in [28]). We independently arrived to similar formulas backed by rigorous proofs, for such performance measures. The performance measures that we cover include

$$\sigma_{i,j} := \lim_{t\to\infty} \frac{\text{Cov}\Big(A_i(t), A_j(t)\Big)}{t}, \tag{1}$$

and the asymptotic variability parameters of flows:

$$\sigma_{i\to j}^2 := \lim_{t\to\infty} \frac{\text{Var}\Big(D_{i,j}(t)\Big)}{t}, \quad \text{and}$$

$$\sigma_{i_1\to j_1, i_2\to j_2} := \lim_{t\to\infty} \frac{\text{Cov}\Big(D_{i_1,j_1}(t), D_{i_2,j_2}(t)\Big)}{t}. \tag{2}$$

The focus and contribution of [28, 29] is on advancing the state of the art in network decomposition approximations and not on exact expressions for asymptotic variability nor on rigorous asymptotic analysis. Hence, in using such results one is left wondering about the meaning and rigor justifying validity of the expressions at hand. Specifically, there remain open questions regarding stability conditions, the usage in diffusion limits and the relationship to asymptotic variance rates.

Our key contribution in the current paper is answering such questions as well as presenting detailed formulas for σ_k^2, $\sigma_{i,j}$, $\sigma_{i\to j}^2$ and $\sigma_{i_1\to j_1,i_2\to j_2}$. Our formulas hold for a wide class of stable networks. However, for concreteness we present our results in the context of stable multi-class queueing networks (of which generalized Jackson networks are a special case). Our main result, Theorem 1, is formulated as a simple functional central limit theorem (FCLT) for the aforementioned processes and also ties the covariance structure of the limiting FCLT processes to system asymptotic variability parameters via uniform integrability. That is, expressions for (1) and (2) and hence for σ_k^2 are rigorously justified.

In dealing with a stable queueing network, this could be viewed as a "fundamental" diffusion limit result similar to some of the results summarized in [10]. To the best of our knowledge within the context of diffusion limits, this fundamental result has been overlooked by previous authors with the exception of [28, 29] that don't focus on rigorous asymptotic results. This is probably due to the fact that much of the exciting research in the field of diffusion approximations of queueing networks in the past three decades has focused on critically loaded networks (cf. [5, 13, 34, 39], as well as many other key references summarized in [10, 17, 31, 37]). The seminal paper [9] does consider diffusion approximations for queueing networks in all regimes (under-loaded, balanced and over-loaded), yet the inter-queue flows are not considered in that paper. Also, in Sect. 4 of the early influential paper [21], Harrison considers flows, but the analysis there is only for the uninterrupted (primitive) processes and not for the true flows $A(\cdot)$ and $D(\cdot)$ as we have here.

As described in our main diffusion result, the asymptotic variability of flows is driven by two components: (i) the variability of the arrival flows; and (ii) the variability resulting from the Bernoulli routing. In stable networks, the variability of queue sizes (related also to service time distributions) does not play a role. Since asymptotic variability of flows only depends on the interplay of the arrival process variability and the Bernoulli routing, we are also motivated to present an alternative way for quantifying the asymptotic variability parameters: *networks with zero service times*. In such networks, jobs that arrive to the network traverse instantaneously through the classes/queues until they depart, and hence the total count of jobs passing on flow $i \to j$ is

$$\sum_k \sum_{\ell=1}^{E_k(t)} N_{i,j|k}(\ell),$$

where the outer sum is over all classes and $N_{i,j|k}(\ell)$ are counts of the number of passes on $i \to j$ for the ℓ'th job arriving exogenously to k. Using elementary calculations, we find the asymptotic variances and covariances of such processes, and prove they

are the same as those originating from the diffusion parameters. It is the zero-service time view which allows us to establish uniform integrability and to relate the diffusion parameters to asymptotic variability parameters.

The structure of the sequel is as follows: In Sect. 2, we summarize our results in a main theorem together with the notation and assumptions of the model. Then, the three sections that follow constitute the proof. In Sect. 3, we present the calculation of the diffusion parameters and diffusion limit. In Sect. 4, we present the alternative view of the network based on zero service times. In Sect. 5, we relate the diffusion parameters to asymptotic variance and establish the required UI. We then follow with Sect. 6 where we present a numerical example, and compare with the innovations method of [28]. For this comparison and for the convenience of readers, we spell out the details of the relevant results from [28]. Readers are encouraged to read Sect. 6 in conjunction with Sect. 2. Closing remarks are in Sect. 7.

2 Model and main result

We consider open multi-class queueing networks (MCQN) operating under arbitrary resource allocation policies and subject to Bernoulli routing. Special cases include generalized Jackson queueing networks as described in [10] as well as many other examples appearing in [4]. Consider a queueing network of J servers serving K classes of customers denoted by $1, \ldots, K$. For each class, there is a unique server $s(k)$ and let $C_j = \{k : s(k) = j\}$ denote the constituency of server j, that is, C_j is the set of all the classes served at server j.

The evolution of the network is driven by the following primitive sequences of random variables: $\{\xi_k(\ell)\}_{\ell=0}^{\infty}$ is the sequence of exogenous inter-arrival times to class k and $\{\eta_k(\ell)\}_{\ell=0}^{\infty}$ is the sequence of service times in class k. The sequence $\{\phi_{i,j}(\ell)\}_{\ell=1}^{\infty}$ of indicator variables determines if the ℓ'th job departing from class i moves to class j (this is indicated via $\phi_{i,j}(\ell) = 1$) where $i \in \{1, \ldots, K\}$ and $j \in \{0, \ldots, K\}$ with $j = 0$ implying the job departing from the network. For any ℓ,

$$\sum_{j=0}^{K} \phi_{i,j}(\ell) = 1. \tag{3}$$

These primitives are assumed to exist on a joint probability space and construct the primitive processes $\big(E, S, \Phi\big)$, as we describe now.

$$E_k(t) = \max\left\{n \geq 0 : \sum_{\ell=0}^{n-1} \xi_k(\ell) \leq t\right\} \quad \text{and} \quad S_k(t) = \max\left\{n \geq 0 \ : \ \sum_{\ell=0}^{n-1} \eta_k(\ell) \leq t\right\},$$

with the convention that summation from 0 to -1 is 0. The counting processes, $E_k(t)$ and $S_k(t)$, represent the number of exogenous arrivals to class k during $[0, t]$ and the number of jobs served during uninterrupted service in class k during $[0, t]$, respectively.

Further, for $\ell = 1, 2, \ldots$, let,

$$\Phi_{i,j}(\ell) = \sum_{\ell'=1}^{\ell} \phi_{i,j}(\ell'),$$

which denotes the number of items routed from class i to class j out of the first ℓ items served at i. Due to (3), for any ℓ,

$$\sum_{j=0}^{K} \Phi_{i,j}(\ell) = \ell, \quad i = 1, \ldots, K. \tag{4}$$

The primitive processes (E, S, Φ) interact to yield the network processes (T, Q, D, A) which we define now. Let $T_k(t)$ denote the work (in units of time) allocated toward serving class k during the time interval $[0, t]$. In general, T is policy dependent as it captures how server effort is allocated among classes. We have for all $j \in \{1, \ldots, J\}$,

$$\sum_{k \in C_j} T_k(t) \leq t.$$

With $T_k(t)$ at hand, the actual number of class k jobs served during $[0, t]$ is $S_k\big(T_k(t)\big)$. Further, composing with Φ we define the *inter-class flows* via

$$D_{i,j}(t) = \Phi_{i,j}\Big(S_i\big(T_i(t)\big)\Big), \quad i = 1, \ldots, K, \;\; j = 0, \ldots, K. \tag{5}$$

Let $Q_k(t)$ denote the number of items of class k at time t in the system (queue or in service). We refer to this number as the queue length, and it satisfies

$$Q_k(t) = A_k(t) - \sum_{j=0}^{K} D_{k,j}(t) + Q_k(0),$$

where the (total) arrival process to class k is

$$A_k(t) = E_k(t) + \sum_{i=1}^{K} D_{i,k}(t). \tag{6}$$

In our exposition, we assume $Q_k(0) = 0$ and thus the queue length equation becomes

$$Q_k(t) = A_k(t) - \sum_{j=0}^{K} D_{k,j}(t). \tag{7}$$

Our results below can be generalized for cases where $Q_k(0)$ is at some fixed positive quantity or is random. For clarity of the exposition, we omit these details.

In the treatment below, the vectors Q, T, E, A and S (and their "bar," "hat" and "tilde" versions as defined below) are treated as K-dimensional column vectors. Further, let Φ and D be K^2-dimensional column vectors with the elements ordered in lexicographic order with the elements $D_{k,0}$ omitted. For example,

$$D = \left(D_{1,1}, \ldots, D_{1,K}, D_{2,1}, \ldots, D_{2,K}, \ldots\ldots\ldots\ldots, D_{K,1}, \ldots, D_{K,K}\right)^{\mathsf{T}}.$$

Probabilistic assumptions

Throughout the paper, we make use of the following assumptions on the network primitives. Without loss of generality assume for some $1 \leq L \leq K$ that only the first L classes have non-null exogenous arrivals, and for the other $K - L$ classes, the exogenous inter-arrival times are infinite (no arrivals). Note that (A3) is not needed for our main result, but is needed to satisfy positive Harris recurrence (stability) results in general.

(A1) $\{\xi_k(\ell)\}_{\ell=0}^{\infty}$ are i.i.d. sequences and mutually independent over all $k = 1, \ldots, L$. Furthermore, independent of inter-arrival times, the sequences $\{\eta_k(\ell)\}_{\ell=0}^{\infty}$ are i.i.d. sequences and mutually independent over all $k = 1, \ldots, K$.

(A2) For all $k = 1, \ldots, K$, we have $0 < \mathbb{E}[(\xi_k(0))^r] < \infty$ and $0 < \mathbb{E}[(\eta_k(0))^r] < \infty$ for $r = 3$. (In Lemma 1, which is not essential to our main result, we assume a slightly stronger version, for some $r > 3$).

(A3) Independent of inter-arrival times and service times, each vector $(\phi_{k,0}(\ell), \ldots, \phi_{k,K}(\ell))^{\mathsf{T}}$ for $k = 1, \ldots, K$ and $\ell \geq 1$ follows a multinomial distribution with a single success and probability vector $(p_{k,0}, p_{k,1}, \ldots, p_{k,K})^{\mathsf{T}}$ with $p_{k,j} \geq 0$, and $p_{k,0} = (1 - \sum_{j=1}^{K} p_{k,j}) \geq 0$. We denote by P the $K \times K$ matrix of $p_{k,j}$, $k, j = 1, \ldots, K$.

Assumption (A1) is standard. Assumption (A2) yields finite moments and is used for diffusion limits and uniform integrability. Assumption (A3) is the standard "Bernoulli routing" assumption implying that each $K + 1$-dimensional vector $(\phi_{k,0}(\ell), \ldots, \phi_{k,K}(\ell))^{\mathsf{T}}$ has a single entry that is 1 and K zero entries.

Since ξ and η have finite third moments, they also have finite first and second moments and we denote

$$\alpha_k = \frac{1}{\mathbb{E}[\xi_k(0)]}, \qquad \text{and} \qquad \mu_k = \frac{1}{\mathbb{E}[\eta_k(0)]}.$$

We also denote by α the vector of α_k and μ the vector of μ_k. Note that later in Sect. 4 we consider a situation where essentially $\mu_k \equiv \infty$ since the service times are zero, yet that section deals with a related model which has processes that can be coupled to the flow processes of the actual model described here.

Structural network assumptions

We assume that the network is *open* and *stabilizable* via the following two assumptions:

(A4) The matrix P has a spectral radius less than 1 that is, $I - P^{\mathsf{T}}$ is non-singular.

We now denote $\lambda = (I - P^{\mathsf{T}})^{-1}\alpha$ and $\lambda_{i,j} := \lambda_i \, p_{i,j}$. Now assume

(A5) $\sum_{k \in \mathcal{C}_j} \frac{\lambda_k}{\mu_k} < 1$ for every server j.

An additional assumption that we make is that policies are *work conserving*. For this, we denote the idle time process of server j via,

$$\mathcal{I}_j(t) = t - \sum_{k \in C_j} T_k(t).$$

Now, the work-conserving assumption is

(A6) $\int_0^t \Big(\sum_{k \in C_j} Q_k(u) \Big) \, d\,\mathcal{I}_j(u) = 0$ for all $t \geq 0$ and all servers j.

Assumption (A4) means that the network is open. Assumption (A5) is used for stability, a concept that we discuss in further detail below. The assumption implies that there is enough capacity in the network. If it is violated, then it is easy to show that the network cannot be stabilized, [4]. As opposed to that under (A5), much research has gone into finding scheduling policies that stabilize the network. For general MCQN such policies exist, see for example [6, 35] and references therein. In the case of generalized Jackson networks (single class meaning that $|C_j| = 1$ for all servers j), under (A5) and (A6) networks are stable, see for example [2].

Scaling limits

For $n = 1, 2, \ldots$ and a function $U(t)$, denote $\overline{U}^n(t) = U(nt)/n$. We say that a fluid limit of U exists if $\lim_{n\to\infty} \overline{U}^n(t) = \overline{U}(t)$ exists uniformly on compact sets (u.o.c) almost surely. Further, when the limit $\overline{U}(t)$ exists, denote

$$\widehat{U}^n(t) = \frac{U(nt) - \overline{U}(nt)}{\sqrt{n}}, \quad n = 1, 2, \ldots. \tag{8}$$

In cases where the above sequence converges weakly on Skorohod J_1 topology to a limiting process, $\widehat{U}(t)$, we denote

$$\widehat{U}^n \Rightarrow \widehat{U}.$$

For discrete time processes replace $U(nt)$ by $U(\lfloor nt \rfloor)$. See [10], Chapter 5 for brief background of weak convergence in the context of queueing networks. An extensive treatment is in [37].

As a consequence of the assumptions (A1) and (A2) (for first moments), the primitive processes satisfy a functional strong law of large numbers (FSLL) yielding fluid limits $\overline{E}_k(t) = \alpha_k t$ and $\overline{S}_k(t) = \mu_k t$. Further, from (A3), $\overline{\Phi}_{i,j}(\ell) = p_{i,j}\ell$.

As a consequence of assumptions (A1) and (A2), the primitive processes satisfy functional central limit theorems (FCLT). Specifically $\widehat{E}_k(t)$ are Brownian motions with diffusion coefficients (also sometimes known as volatility coefficients),

$$v_k = \alpha_k c_k^2, \qquad \text{where} \qquad c_k^2 = \frac{\mathbb{E}[\xi_k(0)^2]}{\mathbb{E}[\xi_k(0)]^2} - 1. \tag{9}$$

Similar diffusion limits exist for the service processes; however, these do not play a role in our limiting results. The routing processes also have diffusion limits due to (A3). We have that

$$\widehat{\Phi}_{k,\cdot}(t) = \Big(\widehat{\Phi}_{k,1}(t), \ldots, \widehat{\Phi}_{k,K}(t)\Big), \quad k = 1, \ldots, K, \tag{10}$$

are K-dimensional Brownian motions with covariance matrices Γ_k, having the i, j'th entry $p_{k,i}(\delta_{i,j} - p_{k,j})$, where $\delta_{i,j}$ is the Kronecker delta; see [9, 37]. All these results are for primitive processes; our theorem deals with scaling limits of the flows.

Scaled queues convergence assumption

We also require an assumption on the sequence of processes $\widehat{Q}^n(t)$. Specifically the assumption is the following process level convergence:

(A7) $\widehat{Q}^n \Rightarrow 0$ as $n \to \infty$, where 0 is the K-dimensional zero process.

In many cases (A7) is not difficult to verify. For example, in stable generalized (single class) Jackson networks it is automatically satisfied due to our primitive network assumptions (A1)–(A6), see Theorem 7.25 in [10]. There, the proof uses the continuous mapping theorem and the oblique reflection mapping (specific to Generalized Jackson networks) to establish $\widehat{Q}^n \Rightarrow 0$. Further in Proposition 8.12 of [10] one sees such a result for a specific multi-class queueing network with 4 nodes and a specific policy.

However, in a more general context, such as the general multi-class queueing networks that we consider, we have not been able to establish (A7) based on first principles and hence we present (A7) as an assumption. Nevertheless, if one assumes or establishes *tightness* (see, e.g., Chapter VI of [24]) of $\widehat{Q}^n$ using first principles and/or other primitive model assumptions, then Lemma 1, presented in the sequel, yields (A7).

Stability using fluid models

We now briefly give an overview of the fluid stability framework, referring to details in the literature for the sake of brevity. For a MCQN and a scheduling policy, we can associate a set of deterministic equations called the *fluid model (equations)*. Such equations are spelled out in detail in [4] page 104, equations (4.50)–(4.55). The fluid

model description in [4] summarizes key ideas from [8, 11, 12] and others. In general, each scheduling policy may induce a different set of equations. Examples are in [4]. A key object in the fluid model equations is the queue fluid limit, $\{Z(t)\}_{t\geq 0}$, a K-dimensional vector (using the notation of [4]).

The concept of *fluid model stability* then requires that there exist some finite time t^* such that if $\sum_{k=1}^{K} Z_k(0) = 1$ then $\sum_{k=1}^{K} Z_k(t) = 0$ for $t \geq t^*$. Much of the literature on MCQN has dealt with proving fluid model stability associated with different networks and scheduling policies, see [4]. A key result in [11] (also summarized in [4]) connects fluid model stability to positive Harris recurrence of the associated Markov process describing the MCQN. In general, the most accepted stability notion of a stochastic MCQN is positive Harris recurrence. Of notable mention are generalized Jackson networks (single class) which are stable under any work-conserving policy and assumptions (A4)–(A5). Our main theorem requires the fluid model of the network to be stable.

Assumptions (A1)–(A3) and an additional technical assumption requiring the interarrival times to be unbounded and spread-out (see for example (1.4) and (1.5) in [11]) can be used to show that a stable fluid model yields positive harris recurrence of the network. In this paper, we don't explicitly require positive Harris recurrence, and hence such an additional technical assumption is not needed.

A related stability notion that we use in the sequel is weak stability. The fluid model is *weakly stable* if when $Z(0) = 0$ then $Z(t) = 0$ for all $t \geq 0$. Weak stability clearly follows from fluid model stability since our networks are time-homogenous.

Main result

We now set up some matrices and vectors used in our main theorem. Use $\mathbf{1}$ to denote the K-dimensional vector of ones and define the $K \times K^2$ matrix $B := \mathbf{1}^{\mathsf{T}} \otimes I$ where $\otimes$ is the Kronecker product and here I is the $K \times K$ identity matrix. Further denote the $K^2 \times K$ matrix,

$$P_c := \begin{bmatrix} P^{\mathsf{T}} e_{1,1} \\ P^{\mathsf{T}} e_{2,2} \\ \vdots \\ P^{\mathsf{T}} e_{K,K} \end{bmatrix},$$

where $e_{i,j}$ is a $K \times K$ matrix with all entries 0 except for the i,j'th entry being 1. Now define the $K \times (K + K^2)$ matrix G and the $K^2 \times (K + K^2)$ matrix H, respectively, as,

$$G := \left[(I - P^{\mathsf{T}})^{-1} \; (I - P^{\mathsf{T}})^{-1} B \right], \tag{11}$$

$$H := \left[P_c (I - P^{\mathsf{T}})^{-1} \; I_{K^2} + P_c (I - P^{\mathsf{T}})^{-1} B \right]. \tag{12}$$

Also define the $(K + K^2) \times (K + K^2)$ covariance matrix for the exogenous arrival processes and the routing processes:

$$\Sigma^{(P)} := \begin{bmatrix} \text{diag}(v_k^2) & & & 0 \\ & \lambda_1 \Gamma_1 & & \\ & & \ddots & \\ 0 & & & \lambda_K \Gamma_K \end{bmatrix}, \tag{13}$$

where $\text{diag}(v_k^2)$ is a diagonal matrix with elements v_k^2. Further, for any $i, j \in \{1, \ldots, K\}$ define the K-dimensional vector $m(i, j)$ as follows:

$$m(i, j) := (I - P)^{-1} e_{i,i} P_{\cdot,j}, \tag{14}$$

where $P_{\cdot,j}$ is the j'th column of P. As further elaborated on in Sect. 4, the k'th entry of the column vector $m(i, j)$ is the expected number of transitions from state i to state j in a Markov chain whose transient component is specified by P and initial state is set to k.

We now present our main result. Relationships to the results of [28] are in Sect. 6.

Theorem 1 *Consider a multi-class queueing network and assume (A1)–(A7) hold. If the fluid model of the network (incorporating the scheduling policy) is stable, then*

(i) The sequences $\widehat{A}^n$ and $\widehat{D}^n$ converge weakly to drift-less Brownian motion processes with covariance matrices,

$$\Sigma^{(A)} := G\, \Sigma^{(P)}\, G^T, \quad \text{and} \quad \Sigma^{(D)} := H\, \Sigma^{(P)}\, H^T, \tag{15}$$

respectively.

(ii) The asymptotic variability parameters, as defined in (1) and (2), can be read off from the diffusion parameters. Namely,

$$\sigma_{i_1 \to j_1, i_2 \to j_2} = \Sigma^{(D)}_{(i_1-1)K + j_1,\ (i_2-1)K + j_2}, \qquad \sigma_{i,j} = \Sigma^{(A)}_{i,j}.$$

(iii) An alternative calculation for the asymptotic variability parameters is

$$\begin{aligned} \sigma_{i_1 \to j_1, i_2 \to j_2} &= m_{j_1}(i_2, j_2) \alpha^T m(i_1, j_1) + m_{j_2}(i_1, j_1) \alpha^T m(i_2, j_2) \\ &\quad + (v^2 - \alpha)^T \big(m(i_1, j_1) \bullet m(i_2, j_2)\big), \end{aligned} \tag{16}$$

$$\sigma^2_{i \to j} = (1 + 2m_j(i, j)) \alpha^T m(i, j) + (v^2 - \alpha)^T \big(m(i, j) \bullet m(i, j)\big), \tag{17}$$

$$\sigma_{i,j} = v_i^2 \sum_{k=1}^{K} m_i(k, j) + v_j^2 \sum_{k=1}^{K} m_j(k, i) + \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sigma_{k_1 \to i, k_2 \to j} \tag{18}$$

where $m_k(i, j)$ is the kth entry of the vector $m(i, j)$ and $(x \bullet y)$ signifies the vector resulting from element-wise product of the vectors x and y.

The proof is structured as follows: (i) is established in Sect. 3. (iii) is established in Sect. 4. (ii) relies on the development of (iii) and is established in Sect. 5. Since the result may appear quite technical, we demonstrate the applicability on a specific network example in Sect. 6.

We also put forward foundations utilizing several results from the literature. Specifically equations (19), (20), (21), and (22), that we obtain now, are used in the proofs.

Using weak stability, and assumptions (A4)–(A6), Theorem 4.1 in [8] ensures $\overline{Q}(t)$ and $\overline{T}(t)$ exist and for any class k,

$$\overline{Q}_k(t) = 0, \tag{19}$$

and

$$\overline{T}_k(t) = \frac{\lambda_k}{\mu_k} t. \tag{20}$$

Using fluid model stability and assumptions (A1) and (A2), Theorem 4.1 (ii) in [12] states that for every initial state x of the associated Markov process of the network (and policy),

$$\lim_{t\to\infty} \mathbb{E}_x \left[Q_k(t)^2 \right] \leq c, \tag{21}$$

for some constant c and any class k.

We now refer to equation (5.18) on page 60 of [20]. We have that under assumptions (A1) and (A2) for the arrival process, for each class k and some $t_0 > 0$,

$$\left\{ \frac{(E_k(t) - \alpha_k t)^2}{t},\ t \geq t_0 \right\} \quad \text{is uniformly integrable.} \tag{22}$$

Finally, of independent interest when one is able to establish tightness of $\widehat{Q}^n$ for a specific model, the following result yields a sufficient condition for $\widehat{Q}^n \Rightarrow 0$ and can be used in place of (A7).

Lemma 1 *Suppose (A1) and (A2) hold, the fluid model is stable, and the sequence of processes $\widehat{Q}^n$ is tight. Suppose further that (A2) holds with $r = 3 + \varepsilon$ for some $\varepsilon > 0$, then $\widehat{Q}^n \Rightarrow 0$ as $n \to \infty$.*

Proof Using the union bound and the Markov inequality, for any $\tilde{\varepsilon} > 0$

$$\mathbb{P}\left(|\widehat{Q}^n(t)| > \tilde{\varepsilon} \right) \leq \sum_{k=1}^{K} \mathbb{P}\left(\frac{Q_k(nt)}{\sqrt{n}} > \frac{\tilde{\varepsilon}}{K} \right) \leq \frac{K^{2+\varepsilon}}{n^{1+\varepsilon/2}\, \tilde{\varepsilon}^{2+\varepsilon}} \sum_{k=1}^{K} \mathbb{E}\left[Q_k(nt)^{2+\varepsilon} \right]. \tag{23}$$

Now, the finite moment convergence result (21) yields $\lim_{t\to\infty} \mathbb{E}_x \left[Q_k(t)^{2+\varepsilon} \right] \leq c$. With the Borel–Cantelli lemma, since the series (in n) summing the probabilities of

the left-hand side of (23) converges, it holds that $|\widehat{Q}^n(t)| \to 0$ as $n \to \infty$ almost surely. Hence for class k and every fixed t $\lim_{n\to\infty} \widehat{Q}^n_k(t) = 0$, almost surely.

To see the convergence of finite-dimensional distributions of $\widehat{Q}^n$, consider $t_1, \ldots, t_\ell$, a finite sequence of time points. From (23) and (21), we have that $\mathbb{P}\left(|\widehat{Q}^n(t)| > \tilde{\varepsilon}\right) \to 0$ as $n \to \infty$. As a simple consequence,

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{Q}^n(t_1)| > \tilde{\varepsilon}, \ldots, |\widehat{Q}^n(t_\ell)| > \tilde{\varepsilon}\right) = 0.$$

Since this is true for any finite sequence $t_1, \ldots, t_\ell$, under the assumption that $\{\widehat{Q}^n : n \geq 1\}$ is tight, from Sect. 3 of Chapter VI in [24], we have the required result. □

3 The diffusion parameters

Using the scaling definition (8), Eqs. (4), (7) and (6) are easily manipulated for all $k = 1, \ldots, K$ and $n \geq 1$ to yield,

$$0 = \sum_{j=0}^{K} \widehat{\Phi}^n_{k,j}(\ell), \quad \ell = 1, 2, \ldots, \tag{24}$$

$$\widehat{Q}^n_k(t) = \widehat{E}^n_k(t) + \sum_{i=1}^{K} \widehat{D}^n_{i,k}(t) - \sum_{j=0}^{K} \widehat{D}^n_{k,j}(t), \quad t \geq 0, \tag{25}$$

$$\widehat{A}^n_k(t) = \widehat{E}^n_k(t) + \sum_{i=1}^{K} \widehat{D}^n_{i,k}(t), \quad t \geq 0. \tag{26}$$

Observe that the property $\overline{T}_k(t) = \frac{\lambda_k}{\mu_k} t$ implies,

$$\lim_{n\to\infty} \overline{D}^n_{i,j}(t) := \overline{D}_{i,j}(t) = \overline{\Phi}_{i,j}(\overline{S}_i(\overline{T}_i(t))) = p_{i,j}\lambda_i t, \quad \text{u.o.c.}. \tag{27}$$

Lemmas 2–5 summarize straightforward algebraic manipulations of these equations. Then, this leads to a simple diffusion limit that follows from Donsker's theorem (see [10], Chapters 5-7 or [17] for background). Techniques similar to those employed here are also in [33], applied to queueing networks that generate their own input. The basic idea is to represent the diffusion scaled processes, $\widehat{D}^n$ and $\widehat{T}^n$ in terms of the following "tilde" processes:

$$\widetilde{\Phi}^n_{i,j}(t) := \widehat{\Phi}^n_{i,j}\left(\overline{S}^n_i\left(\overline{T}^n_i(t)\right)\right), \quad \text{and} \quad \widetilde{S}^n_k(t) := \widehat{S}^n_k(\overline{T}^n_k(t)),$$

which in turn have diffusion limits based on the primitive processes.

Lemma 2 *For $i = 1, \ldots, K$ and $j = 0, \ldots, K$,*

$$\widehat{D}^n_{i,j}(t) = \widetilde{\Phi}^n_{i,j}(t) + p_{i,j}\widetilde{S}^n_i(t) + p_{i,j}\mu_i \widehat{T}^n_i(t). \tag{28}$$

Proof Use $D_{i,j}(nt) = \Phi_{i,j}(S_i(T_i(nt))) = \Phi_{i,j}(n\overline{S}_i^n(\overline{T}_i^n(t)))$ and (27) to get

$$\begin{aligned}\widehat{D}_{i,j}^n(t) &= \frac{\Phi_{i,j}(n\overline{S}_i^n(\overline{T}_i^n(t))) - p_{i,j}\lambda_i nt}{\sqrt{n}} \\ &= \frac{\Phi_{i,j}(n\overline{S}_i^n(\overline{T}_i^n(t))) - p_{i,j}n\overline{S}_i^n(\overline{T}_i^n(t))}{\sqrt{n}} + \frac{p_{i,j}n\overline{S}_i^n(\overline{T}_i^n(t)) - p_{i,j}\mu_i n\overline{T}_i^n(t)}{\sqrt{n}} \\ &\quad + \frac{p_{i,j}\mu_i n\overline{T}_i^n(t) - p_{i,j}\lambda_i nt}{\sqrt{n}}.\end{aligned}$$

Now, (28) follows. □

Denote by M the diagonal matrix with diagonal elements μ_k^{-1}. We now have

Lemma 3 *The diffusion scaled time allocation can be written as:*

$$\widehat{T}^n(t) = M(I - P^\intercal)^{-1}\Big(\widehat{E}^n(t) + B\widetilde{\Phi}^n(t) - \widehat{Q}^n(t)\Big) - M\widetilde{S}^n(t). \tag{29}$$

Proof Substituting (28) into (25), we have

$$\begin{aligned}\widehat{Q}_k^n(t) &= \widehat{E}_k^n(t) + \sum_{i=1}^K \big(\widetilde{\Phi}_{i,k}^n(t) + p_{i,k}\widetilde{S}_i^n(t) + p_{i,k}\mu_i\widehat{T}_i^n(t)\big) \\ &\quad - \sum_{j=0}^K \Big(\widetilde{\Phi}_{k,j}^n(t) + p_{k,j}\widetilde{S}_k^n(t) + p_{k,j}\mu_k\widehat{T}_k^n(t)\Big) \\ &= \widehat{E}_k^n(t) + \sum_{i=1}^K \big(\widetilde{\Phi}_{i,k}^n(t) + p_{i,k}\widetilde{S}_i^n(t) + p_{i,k}\mu_i\widehat{T}_i^n(t)\big) - \widetilde{S}_k^n(t) - \mu_k\widehat{T}_k^n(t) \\ &= \widehat{E}_k^n(t) + \sum_{i=1}^K \widetilde{\Phi}_{i,k}^n(t) - \big(\widetilde{S}_k^n(t) - \sum_{i=1}^K p_{i,k}\widetilde{S}_i^n(t)\big) - \big(\mu_k\widehat{T}_k^n(t) - \sum_{i=1}^K p_{i,k}\mu_k\widehat{T}_i^n(t)\big),\end{aligned}$$

where in the second step we used (24) and $\sum_{j=0}^K p_{i,j} = 1$. In vector/matrix form this reads:

$$\widehat{Q}^n(t) = \widehat{E}^n(t) + B\widetilde{\Phi}^n(t) - (I - P^\intercal)\widetilde{S}^n(t) - (I - P^\intercal)M^{-1}\widehat{T}^n(t).$$

Now (29) follows by multiplying both sides by $M(I - P^\intercal)^{-1}$. □

As a consequence, we now have,

Lemma 4 *The process $\widehat{D}^n(t)$ can be represented as,*

$$\widehat{D}^n(t) = H\begin{bmatrix}\widehat{E}^n(t) \\ \widetilde{\Phi}^n(t)\end{bmatrix} - P_c\,(I - P^\intercal)^{-1}\widehat{Q}^n(t).$$

Proof Equations (28) are

$$\widehat{D}^n(t) = \widetilde{\Phi}^n(t) + P_c\left(\widetilde{S}^n(t) + M^{-1}\widehat{T}^n(t)\right).$$

Substituting (29) in the above, $\widetilde{S}^n(t)$ drops out of the equation, and we obtain,

$$\begin{aligned}\widehat{D}^n(t) = &\left(I_{K^2} + P_c\,(I - P^{\mathsf{T}})^{-1}B\right)\widetilde{\Phi}^n(t)\\ &+P_c\,(I - P^{\mathsf{T}})^{-1}\widehat{E}^n(t) - P_c(I - P^{\mathsf{T}})^{-1}\widehat{Q}^n(t).\end{aligned}$$

□

Observe from Lemma 4 that $\widehat{D}^n$ does not depend directly on $\widehat{S}^n$ but rather through $\widehat{Q}^n$. This means that as $n \to \infty$ the diffusion processes of the service times do not affect the diffusion processes of the flows. We may now represent the analogous result for $\widehat{A}^n$, this time omitting the primitive sequence $\widehat{S}^n$ from the representation.

Lemma 5 *The process $\widehat{A}^n(t)$ can be represented as,*

$$\widehat{A}^n(t) = G\begin{bmatrix}\widehat{E}^n(t)\\ \widetilde{\Phi}^n(t)\end{bmatrix} - B\,P_c(I - P^{\mathsf{T}})^{-1}\widehat{Q}^n(t),$$

Proof We use (26) and the previous lemma:

$$\begin{aligned}\widehat{A}^n(t) &= B\,\widehat{D}^n(t) + \widehat{E}^n(t)\\ &= B\,H\begin{bmatrix}\widehat{E}^n(t)\\ \widetilde{\Phi}^n(t)\end{bmatrix} - B\,P_c(I - P^{\mathsf{T}})^{-1}\widehat{Q}^n(t) + [I_K\ \ 0]\begin{bmatrix}\widehat{E}^n(t)\\ \widetilde{\Phi}^n(t)\end{bmatrix}\\ &= \left(B\,H + [I_K\ \ 0]\right)\begin{bmatrix}\widehat{E}^n(t)\\ \widetilde{\Phi}^n(t)\end{bmatrix} - B\,P_c(I - P^{\mathsf{T}})^{-1}\widehat{Q}^n(t).\end{aligned}$$

Since $B\,P_c = P^{\mathsf{T}}$ and $P^{\mathsf{T}}\,(I - P^{\mathsf{T}})^{-1} = (I - P^{\mathsf{T}})^{-1} - I$,

$$\left(B\,H + [I_K\ \ 0]\right) = \left[B\,P_c(I - P^{\mathsf{T}})^{-1} + I\,\left(I + B\,P_c(I - P^{\mathsf{T}})^{-1}\right)\,B\right] = G.$$

□

We can now establish the diffusion limit in our main theorem.

Proof of Theorem 1 (i): Assumptions (A1) and (A2) imply that for each class k, there are FCLTs for $\widehat{E}^n_k$ with diffusion coefficients as described in (9). Assumption (A3) together with applications of the continuous mapping theorem and (20) imply FCLTs where for each class k, $\widetilde{\Phi}^n_{k,\cdot}(t)$ converges weakly to K-dimensional Brownian motion with covariance matrix $\mu_k \frac{\lambda_k}{\mu_k}\Gamma_k$.

By assumption of mutual independence of primitive processes, as stated in (A1) and (A3), the limiting covariance matrix of

$$\begin{bmatrix} \widehat{E}^n(t) \\ \widetilde{\Phi}^n(t) \end{bmatrix}$$

is $\Sigma^{(P)}$. The result then follows from the representation in Lemmas 4 and 5 and Assumption (A7). □

We note that Lemma 3 can also yield diffusion limits for rate allocations. This appears as (7.89), pp.189 in [10]. In fact, there the authors handle a much wider case in which some queues may be critical and/or over-loaded. This is originally from [9] (6.14), pg 1498. As stated in the Introduction, the diffusion limits for D and A did not appear in [9] and subsequent literature. It is insightful to know that we may also obtain joint diffusion limits for T and D or A, yet we do not pursue this here. Further, handling the case of over-loaded queues does also not pose any additional technical difficulty. The case of critical queues is in general an open question. It was handled in [1] for the single station queue.

4 The zero-service time view

In this section, we refer to the queues as *nodes* to make it clear that there is actually no queueing taking place. For the ℓth customer arriving exogenously first to node k, denote $N_{j|k}(\ell)$ as the number of times that the customer visits node j, and denote $N_{i,j|k}(\ell)$ as the number of times that the customer traverses on the flow $i \to j$. Thus, $N_{j|k}(\ell) = \sum_{i=1}^{K} N_{i,j|k}(\ell)$. Define now

$$\breve{D}_{i,j}(t) := \sum_{k=1}^{K} \sum_{\ell=1}^{E_k(t)} N_{i,j|k}(\ell), \quad \text{and}$$

$$\breve{A}_k(t) := E_k(t) + \sum_{i=1}^{K} \breve{D}_{i,k}(t) = E_k(t) + \sum_{k'=1}^{K} \sum_{\ell=1}^{E_k(t)} N_{k|k'}(\ell).$$

The process $\breve{D}_{i,j}(t)$ is a count of the number of items passing from node i to node j up to time t as if service times are 0. In particular, the ℓ'th customer who arrives at node k by time t ($\ell = 1, \ldots, E_k(t)$) makes an "instantaneous tour" through the nodes, passing $N_{i,j|k}(\ell)$ times on the flow $i \to j$. Similarly, $\breve{A}_k(t)$ is the count of the number of jobs arriving to queue k either exogenously or passing through the network assuming that service times are 0.

By considering both $D(\cdot)$ and $\breve{D}(\cdot)$ on the same probability space, we have that *a.s.*,

$$D_{i,j}(t) \le \breve{D}_{i,j}(t).$$

Denote now,

$$\breve{N}_{i,j}(t) := \breve{D}_{i,j}(t) - D_{i,j}(t).$$

This is the number of future passes on $i \to j$ by customers that are currently in the system (where service times are generally nonzero) at time t. It is obvious from the Markovian nature of the routing that

$$\breve{N}_{i,j}(t) \stackrel{d}{=} \sum_{k=1}^{K} \sum_{\ell=1}^{Q_k(t)} N_{i,j|k}(\ell), \tag{30}$$

where the equality $\stackrel{d}{=}$ is in distribution and for given k,

$$\left\{\left(N_{i,j|k}(\ell),\ i,j \in \{1,\dots,K\}, i \neq j\right),\ \ell = 1,2,\dots\right\},$$

is an i.i.d. sequence (of K^2-dimensional random vectors) whose distribution is induced by a discrete time Markov chain on state space $\{0,1,\dots,K\}$ with transition matrix,

$$\widetilde{P} = \begin{bmatrix} 1 & \mathbf{0}^{\mathsf{T}} \\ \mathbf{1} - P\mathbf{1} & P \end{bmatrix}.$$

The equality in distribution in (30) is a result of the fact that customer routes can be generated a priori and are not state dependent. This presents $\breve{N}_{i,j}(t)$ as a form of a queue of the number of items that are yet to pass via $i \to j$ at time t in the nonzero-service time model.

To construct $N_{i,j|k}(\ell)$, denote by $\{X_n^k\}$ a sequence of states generated by the above Markov chain with $\mathbb{P}(X_0 = k) = 1$ for $k \in \{1,\dots,K\}$. Then, for $i \neq j$, $N_{i,j|k}(\ell)$ has the same distribution as the random variable

$$N_{i,j|k} := \sum_{n=1}^{\infty} \mathbb{I}\{X_{n-1}^k = i, X_n^k = j\},$$

with $\mathbb{I}$ denoting an indicator function. Similarly, $N_{j|k}(\ell)$ is distributed as

$$N_{j|k} := \sum_{n=0}^{\infty} \mathbb{I}\{X_n^k = j\}.$$

Since the queueing network is open (P is sub-stochastic), the only recurrent class in this Markov chain is $\{0\}$ and thus the random variables $N_{i,j|k}$ are proper. It is also a standard exercise to show that they have finite mean and variance.

Denote now

$$\breve{\sigma}_{i,j} := \lim_{t\to\infty} \frac{\mathrm{Cov}\left(\breve{A}_i(t),\ \breve{A}_j(t)\right)}{t}, \quad \text{and} \quad \breve{\sigma}_{i_1\to j_1, i_2\to j_2} := \lim_{t\to\infty} \frac{\mathrm{Cov}\left(\breve{D}_{i_1,j_1}(t),\ \breve{D}_{i_2,j_2}(t)\right)}{t}.$$

As we show now, these variability parameters (of the zero-service time flows) are the same as the variability parameters of the system with queueing:

Proposition 1 *If* (21) *holds, then*

$$\breve{\sigma}^2_{i,j} = \sigma^2_{i,j}, \quad \textit{and} \quad \breve{\sigma}_{i_1\to j_1, i_2\to j_2} = \sigma_{i_1\to j_1, i_2\to j_2}. \tag{31}$$

Proof We present the proof for the asymptotic variability of D, the case of A is similar and is omitted. We have

$$\begin{aligned}
&\left|\mathrm{Cov}\Big(\breve{D}_{i_1,j_1}(t), \breve{D}_{i_2,j_2}(t)\Big) - \mathrm{Cov}\Big(D_{i_1,j_1}(t), D_{i_2,j_2}(t)\Big)\right| \\
&\quad\le \left|\mathrm{Cov}\Big(D_{i_1,j_1}(t), \breve{N}_{i_2,j_2}(t)\Big)\right| + \left|\mathrm{Cov}\Big(D_{i_2,j_2}(t), \breve{N}_{i_1,j_1}(t)\Big)\right| \\
&\qquad + \left|\mathrm{Cov}\Big(\breve{N}_{i_1,j_1}(t), \breve{N}_{i_2,j_2}(t)\Big)\right| \\
&\quad\le \sqrt{\mathrm{Var}\Big(D_{i_1,j_1}(t)\Big)\mathrm{Var}\Big(\breve{N}_{i_2,j_2}(t)\Big)} + \sqrt{\mathrm{Var}\Big(D_{i_2,j_2}(t)\Big)\mathrm{Var}\Big(\breve{N}_{i_1,j_1}(t)\Big)} \\
&\qquad + \sqrt{\mathrm{Var}\Big(\breve{N}_{i_1,j_1}(t)\Big)\mathrm{Var}\Big(\breve{N}_{i_2,j_2}(t)\Big)}.
\end{aligned} \tag{32}$$

For any (i, j), we have that both $\mathrm{Var}(D_{i,j}(t))/t$ and $\mathrm{Var}(\breve{N}_{i,j}(t))$ are bounded from the above uniformly in t; for the latter, this is a consequence of (21). Dividing (32) by t and taking $t \to \infty$ we get the result. □

Note: a version of the above result also exists for the mean rates, λ. In this case, all that is required is finiteness of the first moments of the queues.

We now express the components of $\breve{\sigma}$ in terms of $\mathbb{E}[N_{i,j|k}]$ and $\mathrm{Cov}(N_{i_1,j_1|k}, N_{i_2,j_2|k})$.

Proposition 2 *For all* i_1, i_2, j_1, j_2,

$$\begin{aligned}
\breve{\sigma}_{i_1\to j_1, i_2\to j_2} &= \sum_{k=1}^{K} \alpha_k Cov(N_{i_1,j_1|k}, N_{i_2,j_2|k}) \\
&\quad + \sum_{k=1}^{K} v_k^2 \mathbb{E}[N_{i_1,j_1|k}]\mathbb{E}[N_{i_2,j_2|k}], \\
\breve{\sigma}_{j_1,j_2} &= v_{j_1}^2 \mathbb{E}[N_{j_2|j_1}] + v_{j_2}^2 \mathbb{E}[N_{j_1|j_2}] \\
&\quad + \sum_{k=1}^{K} \alpha_k Cov\big(N_{j_1|k}, N_{j_2|k}\big) \; + \; \sum_{k=1}^{K} v_k^2 \, \mathbb{E}[N_{j_1|k}] \, \mathbb{E}[N_{j_2|k}] \\
&= v_{j_1}^2 \mathbb{E}[N_{j_2|j_1}] + v_{j_2}^2 \mathbb{E}[N_{j_1|j_2}] \; + \; \sum_{i_1=1}^{K} \sum_{i_2=1}^{K} \breve{\sigma}_{i_1\to j_1, i_2\to j_2}.
\end{aligned}$$

Proof We begin with the asymptotic variability of $\breve{D}$, namely $\breve{\sigma}_{i_1\to j_1, i_2\to j_2}$. For illustration, we begin with the variance (even though it is a special case of the covariance calculation that follows). Using the conditional variance rule, we get

$$\begin{aligned}\mathrm{Var}\Big(\breve{D}_{i,j}(t)\Big) &= \sum_{k=1}^{K} \mathrm{Var}\Big(\sum_{\ell=1}^{E_k(t)} N_{i,j|k}(\ell)\Big) \\ &= \sum_{k=1}^{K} \Big(\mathbb{E}[E_k(t)]\,\mathrm{Var}(N_{i,j|k}) + \mathrm{Var}(E_k(t))\,\mathbb{E}[N_{i,j|k}]^2\Big).\end{aligned}$$

Moving onto the covariance, observe that $N_{i_1,j_1|k}(\ell)$ and $N_{i_2,j_2|k'}(\ell)$ are independent whenever $k \neq k'$, hence,

$$\begin{aligned}\mathrm{Cov}\Big(\breve{D}_{i_1,j_1}(t), \breve{D}_{i_2,j_2}(t)\Big) &= \sum_{k=1}^{K} \mathrm{Cov}\Big(\sum_{\ell=1}^{E_k(t)} N_{i_1,j_1|k}(\ell), \sum_{\ell=1}^{E_k(t)} N_{i_2,j_2|k}(\ell)\Big) \\ &= \sum_{k=1}^{K} \Big(\mathbb{E}[E_k(t)]\,\mathrm{Cov}(N_{i_1,j_1|k}, N_{i_2,j_2|k}) \\ &\quad + \mathrm{Var}\big(E_k(t)\big)\,\mathbb{E}[N_{i_1,j_1|k}]\mathbb{E}[N_{i_2,j_2|k}]\Big)\end{aligned}$$

where in the second step we use the conditional covariance rule

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[\mathrm{Cov}(X, Y|Z)\big] + \mathrm{Cov}\big(\mathbb{E}[X|Z], \mathbb{E}[Y|Z]\big).$$

Dividing by t and taking $t \to \infty$ yields the result.

Moving onto the asymptotic variability of $\breve{A}$ (this time treating the variance and the other covariance terms together), we expand and get

$$\begin{aligned}\mathrm{Cov}\big(\breve{A}_{j_1}(t), \breve{A}_{j_2}(t)\big) &= \sum_{i_2=1}^{K} \mathrm{Cov}\big(E_{j_1}(t), \breve{D}_{i_2,j_2}(t)\big) + \sum_{i_1=1}^{K} \mathrm{Cov}\big(E_{j_2}(t), \breve{D}_{i_1,j_1}(t)\big) \\ &\quad + \sum_{i_1=1}^{K}\sum_{i_2=1}^{K} \mathrm{Cov}\big(\breve{D}_{i_1,j_1}(t), \breve{D}_{i_2,j_2}(t)\big) \qquad (33)\end{aligned}$$

To rewrite the first sum on the right-hand side, we can use

$$\begin{aligned}\mathrm{Cov}\big(E_{j_1}(t), \breve{D}_{i_2,j_2}(t)\big) &= \sum_{k=1}^{K} \mathrm{Cov}\big(E_{j_1}(t), \sum_{\ell=1}^{E_k(t)} N_{i_2,j_2|k}(\ell)\big) \\ &= \mathbb{E}[\mathrm{Cov}\big(E_{j_1}(t), \sum_{\ell=1}^{E_{j_1}(t)} N_{i_2,j_2|j_1}(\ell) \;\big|\; E_{j_1}(t)\big)\end{aligned}$$

$$
\begin{aligned}
&+ \mathrm{Cov}\big(\mathbb{E}[E_{j_1}(t) \mid E_{j_1}(t)], \mathbb{E}[\sum_{\ell=1}^{E_{j_1}(t)} N_{i_2,j_2|j_1}(\ell) \mid E_{j_1}(t)]\big) \\
&= \mathrm{Cov}\big(E_{j_1}(t),\, E_{j_1}(t)\,\mathbb{E}[N_{i_2,j_2|j_1}(\ell)]\big) \\
&= \mathrm{Var}\big(E_{j_1}(t)\big)\,\mathbb{E}[N_{i_2,j_2|j_1}(\ell)]
\end{aligned}
$$

with a similar expression holding for the second term, while the third term on the right hand side of (33) can be rewritten using

$$
\begin{aligned}
\mathrm{Cov}\big(\breve{D}_{i_1,j_1}(t), \breve{D}_{i_2,j_2}(t)\big) &= \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} \mathrm{Cov}\left(\sum_{\ell_1=1}^{E_{k_1}(t)} N_{i_1,j_1|k_1}(\ell_1), \sum_{\ell_2=1}^{E_{k_2}(t)} N_{i_2,j_2|k_2}(\ell_2),\right) \\
&= \sum_{k=1}^{K} \mathbb{E}\left[\sum_{\ell=1}^{E_k(t)} \mathrm{Cov}\big(N_{i_1,j_1|k}(\ell), N_{i_2,j_2|k}(\ell)\big)\right] \\
&\quad + \sum_{k=1}^{K} \mathrm{Cov}\big(E_k(t)\,\mathbb{E}[N_{i_1,j_1|k}(\ell)],\ E_k(t)\,\mathbb{E}[N_{i_2,j_2|k}(\ell)]\big) \\
&= \sum_{k=1}^{K} \mathbb{E}[E_k(t)]\mathrm{Cov}\big(N_{i_1,j_1|k}(\ell),\ N_{i_2,j_2|k}(\ell)\big) \\
&\quad + \sum_{k=1}^{K} \mathrm{Var}\big(E_k(t)\big)\,\mathbb{E}[N_{i_1,j_1|k}(\ell)]\,\mathbb{E}[N_{i_2,j_2|k}(\ell)].
\end{aligned}
$$

where we used the independence of different customers in the absence of queuing.

Substituting in (33) and using $\sum_{i=1}^{K} N_{i,j|k}(\ell) = N_{j|k}(\ell)$ we arrive at

$$
\begin{aligned}
\mathrm{Cov}\big(\breve{A}_{j_1}(t), \breve{A}_{j_2}(t)\big) &= \mathrm{Var}\big(E_{j_1}(t)\big)\,\mathbb{E}[N_{j_2|j_1}(\ell)] \\
&\quad + \mathrm{Var}\big(E_{j_2}(t)\big)\,\mathbb{E}[N_{j_1|j_2}(\ell)] \\
&\quad + \sum_{k=1}^{K} \mathbb{E}[E_k(t)]\,\mathrm{Cov}\big(N_{j_1|k}(\ell), N_{j_2|k}(\ell)\big) \\
&\quad + \mathrm{Var}\big(E_k(t)\big)\,\mathbb{E}[N_{j_1|k}(\ell)]\mathbb{E}[N_{j_2|k}(\ell)].
\end{aligned}
$$

Now dividing by t and letting $t \to \infty$, the result is immediate. □

We now represent $\mathbb{E}[N_{i,j|k}]$ and $\mathrm{Cov}(N_{i_1,j_1|k}, N_{i_2,j_2|k})$ in terms of the routing matrix P. It is an elementary application of "first step analysis" to calculate the desired moments (cf. [25] and/or [27]), yet we have not seen this specific calculation elsewhere,

so we spell out the details. Define:

$$m(i,j) := \begin{bmatrix} \mathbb{E}[N_{i,j|1}] \\ \vdots \\ \mathbb{E}[N_{i,j|K}] \end{bmatrix}, \quad m(i_1,j_1,i_2,j_2) := \begin{bmatrix} \mathbb{E}[N_{i_1,j_1|1}\, N_{i_2,j_2|1}] \\ \vdots \\ \mathbb{E}[N_{i_1,j_1|K}\, N_{i_2,j_2|K}] \end{bmatrix},$$
$$c(i_1,j_1,i_2,j_2) := \begin{bmatrix} \mathrm{Cov}\big(N_{i_1,j_1|1}, N_{i_2,j_2|1}\big) \\ \vdots \\ \mathrm{Cov}\big(N_{i_1,j_1|K}, N_{i_2,j_2|K}\big) \end{bmatrix}.$$

Lemma 6 *The definition of $m(i,j)$ in (14) agrees with the above, namely*

$$m(i,j) = (I-P)^{-1} e_{i,i}\, P_{\cdot,j}.$$

Further, let $i_1 \to j_1$ and $i_2 \to j_2$ be distinct flows (i.e., $i_1 \neq i_2$, or $j_1 \neq j_2$, or both), then

$$\begin{aligned} m(i_1,j_1,i_2,j_2) &= m(i_1,j_1) m_{j_1}(i_2,j_2) + m(i_2,j_2) m_{j_2}(i_1,j_1), \\ m(i,j,i,j) &= m(i,j)\big(1+2m_j(i,j)\big), \end{aligned} \tag{34}$$

and thus,

$$\begin{aligned} c(i_1,j_1,i_2,j_2) &= m(i_1,j_1)\, m_{j_1}(i_2,j_2) + m(i_2,j_2)\, m_{j_2}(i_1,j_1) - m(i_1,j_1) \bullet m(i_2,j_2), \\ c(i,j,i,j) &= m(i,j)\big(1+2m_j(i,j)\big) - m(i,j) \bullet m(i,j). \end{aligned} \tag{35}$$

Proof It is well known that $\mathbb{E}[N_{i|k}]$ is the (k,i)th element of $(I-P)^{-1}$, and clearly $\mathbb{E}[N_{i,j|k}] = \mathbb{E}[N_{i|k}]\, p_{i,j}$, from which the first statement follows. For $\mathbb{E}[N_{i_1,j_1|k}\, N_{i_2,j_2|k}]$ we condition on the first transition from the initial node k, as follows (let $i_1 \neq i_2$, and/or $j_1 \neq j_2$).

$$\begin{aligned} \mathbb{E}[N_{i_1,j_1|k}\, N_{i_2,j_2|k}] &= \sum_{k'=1, k' \notin \{j_1,j_2\}}^{K} p_{k,k'}\, \mathbb{E}[N_{i_1,j_1|k'} N_{i_2,j_2|k'}] \\ &\quad + p_{k,j_1} \mathbb{E}[(\delta_{k,i_1} + N_{i_1,j_1|j_1}) N_{i_2,j_2|j_1}] \\ &\quad + p_{k,j_2} \mathbb{E}[N_{i_1,j_1|j_2} (\delta_{k,i_2} + N_{i_2,j_2|j_2})] \\ &= \sum_{k'=1}^{K} p_{k,k'}\, \mathbb{E}[N_{i_1,j_1|k'}\, N_{i_2,j_2|k'}] \\ &\quad + p_{k,j_1} \delta_{k,i_1} \mathbb{E}[N_{i_2,j_2|j_1}] + p_{k,j_2} \delta_{k,i_2} \mathbb{E}[N_{i_1,j_2|j_2}]. \end{aligned} \tag{36}$$

Equation (36) can be represented as

$$m(i_1,j_1,i_2,j_2) = P\, m(i_1,j_1,i_2,j_2) + e_{i_1,i_1} P_{\cdot,j_1}\, m_{j_1}(i_2,j_2) + e_{i_2,i_2} P_{\cdot,j_2}\, m_{j_2}(i_1,j_1).$$

or rearranged to

$$m(i_1, j_1, i_2, j_2) = (I - P)^{-1} \left(e_{i_1,i_1} P_{\cdot,j_1} m_{j_1}(i_2, j_2) + e_{i_2,i_2} P_{\cdot,j_2} m_{j_2}(i_1, j_1)\right),$$

which yields (34). In a similar way, we can show that

$$\mathbb{E}[N^2_{i,j|k}] = \sum_{k'=1}^{K} p_{k,k'} \mathbb{E}[N^2_{i,j|k'}] + p_{k,j}\delta_{k,i}\left(1 + 2\mathbb{E}[N_{i,j|j}]\right),$$

which gives

$$m(i, j, i, j) = (I - P)^{-1} e_{i,i} P_{\cdot,j}\left(1 + 2m_j(i, j)\right).$$

□

Proof of Theorem 1 (iii): Proposition 1 indicates that under property (21), the variability parameters are the same as those of the zero-service time processes, and (21) follows from the theorem assumptions. Now, the combination of Proposition 2 and Lemma 6 yields the result. □

5 Asymptotic variance and uniform integrability

As stated at onset, our original goal is to obtain expressions for σ_{k_1,k_2} and $\sigma_{i_1\to j_1,i_2\to j_2}$. As we state in Theorem 1 (ii) these can now be read off from the matrices $\Sigma^{(A)}$ and $\Sigma^{(D)}$, respectively. The presentation in this section is for the $\sigma_{i_1\to j_1,i_2\to j_2}$ terms; analogous results for the terms associated with $A(\cdot)$ can be proved in the exact same manner.

Proving Theorem 1 (ii) requires establishing suitable uniform integrability (UI) conditions for the following families:

$$\mathcal{D}^{(1)}_{i,j} = \left\{ \frac{D_{i,j}(t) - \lambda_{i,j}t}{\sqrt{t}},\ t \geq t_0 \right\},$$

$$\mathcal{D}^{(2)}_{i,j} = \left\{ \frac{\left(D_{i,j}(t) - \lambda_{i,j}t\right)^2}{t},\ t \geq t_0 \right\},$$

$$\mathcal{D}_{(i_1,j_1),(i_2,j_2)} = \left\{ \frac{\left(D_{i_1,j_1}(t) - \lambda_{i_1,j_1}t\right)\left(D_{i_2,j_2}(t) - \lambda_{i_2,j_2}t\right)}{t},\ t \geq t_0 \right\},$$

where $t_0 > 0$ is arbitrary. Note that while each of the families $\mathcal{D}^{(2)}_{i,j}$ is a special case of $\mathcal{D}_{(i_1,j_1),(i_2,j_2)}$, we treat it separately in this section for clarity. See for example [19] for properties of UI sequences and families, and relations to weak convergence.

The following proposition relates the diffusion parameters to the asymptotic variance parameters.

Proposition 3 *If $\mathcal{D}^{(1)}_{i,j}$ and $\mathcal{D}^{(2)}_{i,j}$ are UI, then*

$$\sigma^2_{i\to j} = \Sigma^{(D)}_{(i-1)K+j,\,(i-1)K+j}\,.$$

If $\mathcal{D}^{(1)}_{i,j}$ and $\mathcal{D}_{(i_1,j_1),(i_2,j_2)}$ are UI, then

$$\sigma_{i_1\to j_1, i_2\to j_2} = \Sigma^{(D)}_{(i_1-1)K+j_1,(i_2-1)K+j_2}\,.$$

Proof By the projection map at time $t = 1$ (cf. [37]), we have the convergence in distribution:

$$\frac{D_{i,j}(t) - \lambda_{i,j}t}{\sqrt{t}} \Rightarrow \widehat{D}_{i,j}(1).$$

Further, using the continuous mapping theorem, we obtain

$$\frac{\big(D_{i,j}(t) - \lambda_{i,j}t\big)^2}{t} \Rightarrow \big(\widehat{D}_{i,j}(1)\big)^2.$$

Similarly we have the convergence in distribution on $\mathbb{R}^2$:

$$\Big[\frac{D_{i_1,j_1}(t) - \lambda_{i_1,j_1}t}{\sqrt{t}},\ \frac{D_{i_2,j_2}(t) - \lambda_{i_2,j_2}t}{\sqrt{t}}\Big] \Rightarrow \Big[\widehat{D}_{i_1,j_1}(1),\ \widehat{D}_{i_2,j_2}(1)\Big],$$

and thus using the continuous mapping theorem,

$$\frac{D_{i_1,j_1}(t) - \lambda_{i_1,j_1}t}{\sqrt{t}} \cdot \frac{D_{i_2,j_2}(t) - \lambda_{i_2,j_2}t}{\sqrt{t}} \Rightarrow \widehat{D}_{i_1,j_1}(1) \cdot \widehat{D}_{i_2,j_2}(1).$$

Under the UI conditions established below the above weak convergences in distribution imply that

$$\lim_{t\to\infty} \mathbb{E}\Big[\frac{D_{i,j}(t) - \lambda_{i,j}t}{\sqrt{t}}\Big] = \mathbb{E}\big[\widehat{D}_{i,j}(1)\big],$$
$$\lim_{t\to\infty} \mathbb{E}\Big[\frac{(D_{i,j}(t) - \lambda_{i,j}t)^2}{t}\Big] = \mathbb{E}\big[\big(\widehat{D}_{i,j}(1)\big)^2\big],$$

as well as

$$\lim_{t\to\infty} \mathbb{E}\Big[\frac{\big(D_{i_1,j_1}(t) - \lambda_{i_1,j_1}t\big)}{\sqrt{t}} \cdot \frac{\big(D_{i_2,j_2}(t) - \lambda_{i_2,j_2}t\big)}{\sqrt{t}}\Big] = \mathbb{E}[\widehat{D}_{i_1,j_1}(1) \cdot \widehat{D}_{i_2,j_2}(1)].$$

Combining this implies that

$$\begin{aligned}\sigma^2_{i\to j} &= \lim_{t\to\infty}\frac{\mathrm{Var}\big(D_{i,j}(t)\big)}{t} = \lim_{t\to\infty}\frac{\mathrm{Var}\big(D_{i,j}(t)-\lambda_{i,j}t\big)}{t}\\ &= \lim_{t\to\infty}\frac{\mathbb{E}[(D_{i,j}(t)-\lambda_{i,j}t)^2]}{t} - \Big(\lim_{t\to\infty}\frac{\mathbb{E}[D_{i,j}(t)-\lambda_{i,j}t]}{\sqrt{t}}\Big)^2\\ &= \mathbb{E}[\big(\widehat{D}_{i,j}(1)\big)^2] - \big(\mathbb{E}[\widehat{D}_{i,j}(1)]\big)^2 = \mathrm{Var}\big(\widehat{D}_{i,j}(1)\big) = \Sigma^{(D)}_{(i-1)K+j,\,(i-1)K+j}.\end{aligned}$$

Similarly,

$$\begin{aligned}\sigma_{i_1\to j_1,i_2\to j_2} &= \lim_{t\to\infty}\frac{\mathrm{Cov}\big(D_{i_1,j_1}(t),\,D_{i_2,j_2}(t)\big)}{t}\\ &= \lim_{t\to\infty}\frac{\mathrm{Cov}\big(D_{i_1,j_1}(t)-\lambda_{i_1,j_1}t,\,D_{i_2,j_2}(t)-\lambda_{i_2,j_2}t\big)}{t}\\ &= \lim_{t\to\infty}\frac{\mathbb{E}[(D_{i_1,j_1}(t)-\lambda_{i_1,j_1}t)(D_{i_2,j_2}(t)-\lambda_{i_2,j_2}t)]}{t}\\ &\quad -\Big(\lim_{t\to\infty}\frac{\mathbb{E}[D_{i_1,j_1}(t)-\lambda_{i_1,j_1}t]}{\sqrt{t}}\Big)\Big(\lim_{t\to\infty}\frac{\mathbb{E}[D_{i_2,j_2}(t)-\lambda_{i_2,j_2}t]}{\sqrt{t}}\Big)\\ &= \mathrm{Cov}\Big(\widehat{D}_{i_1,j_1}(1),\ \widehat{D}_{i_2,j_2}(1)\Big) = \Sigma^{(D)}_{(i_1-1)K+j_1,\,(i_2-1)K+j_2}.\end{aligned}$$

□

In establishing the UI, we make use of the following useful inequality: For $r > 1$ and arbitrary real values $z_1, \ldots, z_K$,

$$\Big|\sum_{k=1}^{K} z_k\Big|^r \le K^{r-1}\sum_{k=1}^{K}|z_k|^r, \tag{37}$$

which is a simple consequence of Jensen's inequality. We now establish the required UI.

Proposition 4 *If* (20) *and* (22) *hold, then the families of random variables* $\mathcal{D}^{(1)}_{i,j}$, $\mathcal{D}^{(2)}_{i,j}$ *and* $\mathcal{D}_{(i_1,j_1),(i_2,j_2)}$ *are UI.*

Proof We first note that UI of $\mathcal{D}^{(2)}_{i,j}$ implies UI of the other two types of families as well, due to Theorem 4.7 (with $p = q = 2$) in Chapter 5 of [19]. To establish UI of $\mathcal{D}^{(2)}_{i,j}$, recall from the previous section the representation $D_{i,j}(t) = \breve{D}_{i,j}(t) - \breve{N}_{i,j}(t)$, where $\breve{D}_{i,j}(t)$ is the number of instantaneous passes on flow $i \to j$, and $\breve{N}_{i,j}(t)$ is the number of future passes on that flow. By applying (37) for $r = 1$ and $r = 2$,

respectively, we have

$$\left|\frac{D_{i,j}(t)-\lambda_{i,j}t}{\sqrt{t}}\right| \le \left|\frac{\breve{D}_{i,j}(t)-\lambda_{i,j}t}{\sqrt{t}}\right| + \left|\frac{\breve{N}_{i,j}(t)}{\sqrt{t}}\right|,$$
$$\left|\frac{\big(D_{i,j}(t)-\lambda_{i,j}t\big)^2}{t}\right| \le 2\Big(\left|\frac{\breve{D}_{i,j}(t)-\lambda_{i,j}t}{\sqrt{t}}\right|^2 + \left|\frac{\breve{N}_{i,j}(t)}{\sqrt{t}}\right|^2\Big).$$

It thus suffices to show that

$$\breve{\mathcal{D}}^{(2)}_{i,j} := \left\{\frac{\big(\breve{D}_{i,j}(t)-\lambda_{i,j}t\big)^2}{t},\ t\ge t_0\right\}, \quad \text{and} \quad \breve{\mathcal{N}}^{(2)}_{i,j} := \left\{\frac{\big(\breve{N}_{i,j}(t)\big)^2}{t},\ t\ge t_0\right\},$$

are UI.

To see $\breve{\mathcal{D}}^{(2)}_{i,j}$ is UI, it is useful to denote

$$\breve{D}_{i,j|k}(t) := \sum_{\ell=1}^{E_k(t)} N_{i,j|k}(\ell) \quad \text{and} \quad \lambda_{i,j|k} := \alpha_k \mathbb{E}[N_{i,j|k}].$$

Note that since, $\breve{D}_{i,j}(t) = \sum_{k=1}^K \breve{D}_{i,j|k}(t)$, we have $\sum_{k=1}^K \lambda_{i,j|k} = \lambda_{i,j}$. We now get

$$\begin{aligned}\left|\frac{\big(\breve{D}_{i,j}(t)-\lambda_{i,j}t\big)^2}{t}\right| &= \left|\frac{\big(\sum_{k=1}^K \breve{D}_{i,j|k}(t) - (\sum_{k=1}^K \lambda_{i,j|k})t\big)^2}{t}\right| \\ &= \left(\frac{\big|\sum_{k=1}^K \big(\breve{D}_{i,j|k}(t)-\lambda_{i,j|k}t\big)\big|}{\sqrt{t}}\right)^2 \\ &\le K\sum_{k=1}^K \left|\frac{\breve{D}_{i,j|k}(t)-\lambda_{i,j|k}t}{\sqrt{t}}\right|^2.\end{aligned}$$

In the above, we again used (37) with $r=2$. We now need to show that the families

$$\left\{\frac{\big(\breve{D}_{i,j|k}(t)-\lambda_{i,j|k}t\big)^2}{t},\ t\ge t_0\right\},$$

are UI:

$$\begin{aligned}&\frac{\big(\breve{D}_{i,j|k}(t)-\lambda_{i,j|k}t\big)^2}{t} \\ &= \frac{\big(\sum_{\ell=1}^{E_k(t)} N_{i,j|k}(\ell)-\lambda_{i,j|k}t\big)^2}{t} \\ &= \frac{\big(\sum_{\ell=1}^{E_k(t)+1} N_{i,j|k}(\ell)-\lambda_{i,j|k}t - N_{i,j|k}(E_k(t)+1)\big)^2}{t}\end{aligned}$$

$$= \frac{\left(\sum_{\ell=1}^{E_k(t)+1}\left(N_{i,j|k}(\ell)-\frac{\lambda_{i,j|k}}{\alpha_k}\right)+(E_k(t)+1)\frac{\lambda_{i,j|k}}{\alpha_k}-\lambda_{i,j|k}t-N_{i,j|k}(E_k(t)+1)\right)^2}{t}$$

$$= \frac{\left(\sum_{\ell=1}^{E_k(t)+1}\left(N_{i,j|k}(\ell)-\frac{\lambda_{i,j|k}}{\alpha_k}\right)+((E_k(t)+1)-\alpha_k t)\frac{\lambda_{i,j|k}}{\alpha_k}-N_{i,j|k}(E_k(t)+1)\right)^2}{t}$$

$$\le 3\left(\frac{\left(\sum_{\ell=1}^{E_k(t)+1}\left(N_{i,j|k}(\ell)-\frac{\lambda_{i,j|k}}{\alpha_k}\right)\right)^2}{t}+\frac{\left(((E_k(t)+1)-\alpha_k t)\frac{\lambda_{i,j|k}}{\alpha_k}\right)^2}{t}+\frac{\left(N_{i,j|k}(E_k(t)+1)\right)^2}{t}\right).$$

The first term is a stopped random walk with zero mean increments where $E_k(t)+1$ is UI by (22). Thus, due to Theorems 6.1–6.3 in [20], the first term is UI. The second term is UI again by (22). The third term is obviously UI since the family $N_{i,j|k}(\cdot)$ is i.i.d.

To show that $\breve{\mathcal{N}}_{i,j}^{(2)}$ is UI, we need to show that the second moment of $\breve{N}_{i,j}(t)/\sqrt{t}$ converges (to zero). This approach is due to Remark 5.4 in Chapter 5 of [19]. Define $\breve{N}_{i,j|k}^{Q}(t) := \sum_{\ell=1}^{Q_k(t)} N_{i,j|k}(\ell)$, where $Q_k(t)$ is the queue length at node k at time t. Then, the expectation and variance of the random sums $\breve{N}_{i,j|k}^{Q}(t)$, and hence also (by (30)) of $\breve{\mathcal{N}}_{i,j}(t)$, can be expressed in the expectations and variances of $Q_k(t)$ and $N_{i,j|k}(\ell)$, all of which are $O(1)$ by (20). Thus, the result follows. □

Proof of Theorem 1 (ii): Proposition 4 relies on (20) and (22) which follow from the assumptions of the theorem. Proposition 4 then establishes UI of the families needed for Proposition 3 which exactly states (ii). □

6 Numerical example

Consider the 6-node network illustrated in Figure 1 with parameters,

$$P = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mu = \begin{bmatrix} 8.25 \\ 8.25 \\ 5 \\ 8.25 \\ 5 \\ 5 \end{bmatrix},$$

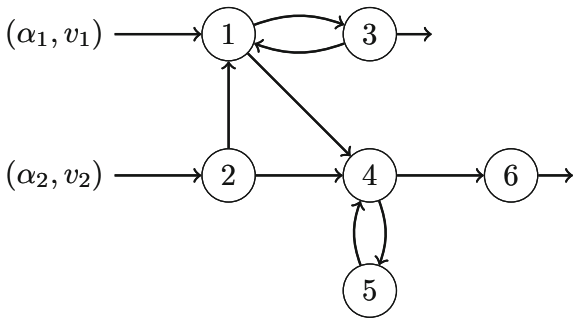

Fig. 1 Example network

Table 1 Covariance values of flows $i \to j$

$i \backslash j$	1	2	3	4	5	6
1	0	0	32/9	20/9	0	0
2	3/2	0	0	3/2	0	0
3	31/18	0	0	0	0	0
4	0	0	0	0	199/18	55/18
5	0	0	0	199/18	0	0
6	0	0	0	0	0	0

$$\alpha = [1 \;\; 4 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^{\mathsf{T}}, \quad \text{and} \quad v^2 = [2 \;\; 2 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^{\mathsf{T}}. \tag{38}$$

For this network,

$$\lambda = (I - P^{\mathsf{T}})^{-1}\alpha = \begin{bmatrix} 4 & 4 & 2 & 8 & 4 & 4 \end{bmatrix}^{\mathsf{T}} < \mu.$$

Hence, if considered as a single-class network, assumptions (A1)–(A7) hold and the network is stabilized by any work-conserving policy. Note that besides verification of the above inequality, the values of μ do not play a further role in the calculation of the variability parameters. Nevertheless, we use them in a simulated example below.

It is now a straightforward matter to use (15) (or alternatively (16)–(18)) from our main theorem to obtain variability parameters. Note that in this process, the only matrix that requires inversion is $(I - P^{\mathsf{T}})$. The rest of the calculations follow from matrix composition, addition and multiplication operations.

The resulting matrix $\Sigma^{(D)}$ is of dimension 36×36. We present the diagonals of this matrix (which are $\sigma^2_{i\to j}$) in Table 1.

As a further illustration, we present a few selected non-diagonal elements of $\Sigma^{(D)}$:

$$\sigma_{2\to 1,2\to 4} = -1/2, \quad \sigma_{4\to 5,5\to 4} = 199/18, \quad \sigma_{1\to 3,4\to 6} = 5/9, \quad \sigma_{1\to 3,2\to 4} = -1/3.$$

In discussing these values, it is good to consider the *asymptotic correlation coefficient*:

$$r_{i_1\to i_2, j_1\to j_2} := \frac{\sigma_{i_1\to i_2, j_1\to j_2}}{\sqrt{\sigma^2_{i_1\to i_2}\sigma^2_{j_1\to j_2}}}.$$

For these selected flow pairs, it evaluates to

$$r_{2\to 1,2\to 4} = -\frac{1}{3}, \quad r_{4\to 5,5\to 4} = 1,$$
$$r_{1\to 3,4\to 6} \approx 0.16856, \quad r_{1\to 3,2\to 4} \approx -0.14434.$$

The first two values are easily explained in our example, the other two are not. For $r_{2\to 1,2\to 4}$, consider the Bernoulli splitting at the output of queue 2 and the fact there is no feedback to this queue. Recall that in this case $\sigma_{2\to 1,2\to 4} = (v_2^2 - \alpha_2)/4$ for $v_2^2 = 2, \alpha_2 = 4$. In this case, the asymptotic correlation coefficient is $(v_2^2 - \alpha_2)/(v_2^2 + \alpha_2)$.

In considering $r_{4\to 5,5\to 4}$, observe that there is no random routing in this part of the network: All jobs that enter 5 come from 4 and then return to 5.

We are not aware of an "easy" explanation of the values of $r_{1\to 3,4\to 6}$ and $r_{1\to 3,2\to 4}$. It is insightful to see that as in this case, some correlations between flows are positive while others are negative. We do not know of an a priori way of finding out the sign of these correlations without using our main result. In fact, evaluating $\Sigma^{(D)}$ with v_2 as free variable, we get

$$r_{1\to 3,2\to 4} = \frac{v_2^2 - 4}{\sqrt{(v_2^2+4)(v_2^2+30)}}.$$

We thus see that the sign of the correlation between those two flows depends on the variability of the arrival process into 2. Observe that in the asymptotically uncorrelated case (i.e., when $v_2 = 4$),

$$\lim_{t\to\infty} \frac{\mathrm{Var}\big(E_2(t)\big)}{\mathbb{E}[E_2(t)]} = 1,$$

as is for a Poisson process. This is consistent with the fact that in the case of a classic Jackson network (Poisson arrival process and exponential processing times) case, since node 2 has no feedback its output is a Poisson process and splitting of departures from node 2 results in two independent Poisson flows, $2 \to 1$ and $2 \to 4$. The first of these flows affects $1 \to 3$ but not the second. Hence, in such a case it is expected that $r_{1\to 3,2\to 4} = 0$.

Arrivals to individual queues

Moving onto arrival processes into individual queues, application of our main result yields

$$\Sigma^{(A)} = \begin{bmatrix} 68/9 & 4/3 & 40/9 & 44/9 & 22/9 & 22/9 \\ & 2 & 2/3 & 10/3 & 5/3 & 5/3 \\ & & 32/9 & 10/9 & 5/9 & 5/9 \\ & & & 182/9 & 127/9 & 55/9 \\ & & & & 199/18 & 55/18 \\ & & & & & 55/18 \end{bmatrix}. \tag{39}$$

Observe that $\sigma_2^2 = 2$ as expected since there are only exogenous arrivals to this queue. Further since all jobs that pass through queue 5 eventually also pass through queue 6, we have

$$\sigma_{k,5} = \sigma_{k,6}, \quad k = 1, 2, 3, 6.$$

It is the diagonal elements of $\Sigma^{(A)}$ that may be useful for network decomposition approximations (which we do not explore further in this paper). Normalizing the

diagonals by λ, we get

$$c^2 = \left[\, 1.89 \;\; 0.5 \;\; 1.78 \;\; 2.53 \;\; 2.76 \;\; 0.76 \,\right]^{\mathsf{T}}. \tag{40}$$

Comparison with the innovations method

To compare our method with the *innovations method* of [28] (see also [29]), we now describe how to compute the asymptotic covariance terms of arrival processes, and the asymptotic variability parameters of flows using the innovations method. Toward that end, we outline a step-by-step procedure to establish equation (42) of [28] under the assumption that all the service times are zero; that is, $\rho_i = 0$, for all $i = 1, 2, \ldots, K$ (using the notation of [28]).

Let $E_i^{(0)}(t) := E_i(t) - \mathbb{E}[E_i(t)]$ be the normalized arrival process into the queue i. Similarly, define $A_i^{(0)}(t)$ and $D_{i,j}^{(0)}(t)$ using $A_i(t)$ and $D_{i,j}(t)$, respectively. Further, let $X(t)$ and $Y(t)$ be two $K^2 + K$-dimensional column vectors defined by

$$X(t) = \begin{bmatrix} E^{(0)}(t) \\ \zeta(t) \end{bmatrix}, \qquad \text{and} \qquad Y(t) = \begin{bmatrix} A^{(0)}(t) \\ D^{(0)}(t) \end{bmatrix},$$

where

$$\begin{aligned} E^{(0)}(t) &= \left[E_1^{(0)}(t), \ldots, E_K^{(0)}(t) \right]^{\mathsf{T}}, \\ A^{(0)}(t) &= \left[A_1^{(0)}(t), \ldots, A_K^{(0)}(t) \right]^{\mathsf{T}}, \\ D^{(0)}(t) &= \left[D_{1,1}^{(0)}(t), \ldots, D_{1,K}^{(0)}(t), D_{2,1}^{(0)}(t), \ldots, D_{K,1}^{(0)}(t), \ldots, D_{K,K}^{(0)}(t) \right]^{\mathsf{T}}, \end{aligned}$$

and $\zeta(t) = \left[\zeta_{1,1}(t), \ldots, \zeta_{1,K}(t), \zeta_{2,1}(t), \ldots, \zeta_{K,1}(t), \ldots, \zeta_{K,K}(t)\right]^{\mathsf{T}}$ is a column vector of so-called innovation processes assumed to have the following properties:

$$\begin{aligned} \mathbb{E}[\zeta_{i,j}(t)] &= 0, \\ \mathrm{Cov}\left(\zeta_{i,j},\, \zeta_{i,k}\right) &= p_{i,j}(\delta_{j,k} - p_{i,k})\mathbb{E}\left[A_i(t)\right], \\ \mathrm{Cov}\left(A_i(t),\, \zeta_{i,j}(t)\right) &= 0, \\ \mathrm{Cov}\left(E_i(t),\, \zeta_{j,k}(t)\right) &= 0 \;\; \text{for all} \;\; i, j \text{ and } k, \\ \mathrm{Cov}\left(\zeta_{i_1,j_1},\, \zeta_{i_2,j_2}\right) &= 0 \;\; \text{for} \;\; i_1 \neq i_2, \;\; \text{and for all } j_1, j_2. \end{aligned}$$

See Sect. 4 of [28] for more details on the innovations.

With these processes and assumptions at hand, [28] presents the following (adapted here to our notation): Let

$$F = \begin{bmatrix} 0 & B \\ P_c & 0 \end{bmatrix},$$

Now using $\rho_i = 0$, for all $i = 1, 2, \ldots, K$, in [28], Eq. (35) of [28] can be re-expressed as

$$Y(t) = F\,Y(t) + X(t),$$

or equivalently, $Y(t) = (I - F)^{-1}\,X(t)$. Non-singularity of $I - F$ follows from that of $I - P^{\mathsf{T}}$. To see this, first note that

$$I - F = \begin{bmatrix} I_K & -B \\ -P_c & I_{K^2} \end{bmatrix}.$$

Using *Banachiewicz inversion formula* for the inverse of a partitioned matrix (see for example [40]), we have

$$\begin{aligned}(I - F)^{-1} &= \begin{bmatrix} (I_K - BP_c)^{-1} & (I_K - BP_c)^{-1}B \\ P_c(I_K - BP_c)^{-1} & I_{K^2} + P_c(I_K - BP_c)^{-1}B \end{bmatrix} \\ &= \begin{bmatrix} (I_K - P^{\mathsf{T}})^{-1} & (I_K - P^{\mathsf{T}})^{-1}B \\ P_c(I_K - P^{\mathsf{T}})^{-1} & I_{K^2} + P_c(I_K - P^{\mathsf{T}})^{-1}B \end{bmatrix} \\ &= \begin{bmatrix} G \\ H \end{bmatrix},\end{aligned}$$

where the second equality follows from the fact that $B\,P_c = P^{\mathsf{T}}$ and the last equality follows from the definitions of G and H, which are given by (11) and (12), respectively. Hence, the claim that $I - F$ is non-singular follows from the assumption that $I - P^{\mathsf{T}}$ is non-singular.

Further, we have that $\mathbb{E}\left[Y(t)\,Y(t)^{\mathsf{T}}\right] = \begin{bmatrix} G \\ H \end{bmatrix} \mathbb{E}\left[X(t)\,X(t)^{\mathsf{T}}\right] \left[G^{\mathsf{T}}\ H^{\mathsf{T}}\right]$, and as a consequence,

$$\widetilde{\Sigma} := \lim_{t\to\infty} \frac{\mathbb{E}\left[Y(t)\,Y(t)^{\mathsf{T}}\right]}{t} = \begin{bmatrix} G \\ H \end{bmatrix} \left(\lim_{t\to\infty} \frac{\mathbb{E}\left[X(t)\,X(t)^{\mathsf{T}}\right]}{t} \right) \left[G^{\mathsf{T}}\ H^{\mathsf{T}}\right]$$

which is equivalent to equation (42) of [28]. Note that, from the above properties of the innovations,

$$\begin{aligned}\lim_{t\to\infty} \frac{\mathbb{E}\left[X(t)\,X(t)^{\mathsf{T}}\right]}{t} &= \begin{bmatrix} \lim_{t\to\infty} \frac{\mathbb{E}\left[E^{(0)}(t)\,E^{(0)}(t)^{\mathsf{T}}\right]}{t} & 0 \\ 0 & \lim_{t\to\infty} \frac{\mathbb{E}\left[\zeta(t)\,\zeta(t)^{\mathsf{T}}\right]}{t} \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{diag}(v^2) & 0 \\ 0 & \lim_{t\to\infty} \frac{\mathbb{E}\left[\zeta(t)\,\zeta(t)^{\mathsf{T}}\right]}{t} \end{bmatrix} \\ &= \Sigma^{(P)},\end{aligned}$$

where $\Sigma^{(P)}$ is defined by (13) and the last equality follows from the above properties of the innovation processes and the fact that $\lim_{t\to\infty} \frac{\mathbb{E}[A_i(t)]}{t} = \lambda_i$ for all i. Therefore,

$$\widetilde{\Sigma} = \begin{bmatrix} G \\ H \end{bmatrix} \Sigma^{(P)} \begin{bmatrix} G^\intercal & H^\intercal \end{bmatrix} = \begin{bmatrix} G\,\Sigma^{(P)}\,G^\intercal & G\,\Sigma^{(P)}\,H^\intercal \\ H\,\Sigma^{(P)}\,G^\intercal & H\,\Sigma^{(P)}\,H^\intercal \end{bmatrix},$$

and thus $\sigma_{k_1,k_2} = \widetilde{\Sigma}_{k_1,k_2}$ and $\sigma_{i_1\to j_1,i_2\to j_2} = \widetilde{\Sigma}_{i_1*K+j_1,i_2*K+j_2}$. Observe that the innovations method and the results of Theorem 1 (i) provide essentially the same expressions.

Finally, observe that the virtue of the innovations method in [28] is that it also allows (and focuses on) cases where ρ_i's (in that paper) are not 0. However, in such cases, all calculations are merely a heuristic. Further, as that was not the focus of [28], the limits and model assumptions in that paper are not rigorously justified as in the current paper.

Simulation results

To further illustrate our result and explore the effect of different policies and constraints on the variance of flows, we carried out a Monte Carlo simulation of the example network.

In the simulation, we set the service distributions of queue k to be distributed as a sum of two i.i.d. exponential random variables, each with mean $(2\mu_k)^{-1}$. This results in a so-called Erlang 2 distribution (having a squared coefficient of variation of $1/2$) with mean μ_k^{-1}.

The arrival process, $E_1(\cdot)$, is the more variable of the two arrival processes. It is taken to be a renewal process of inter-arrival times that are distributed as a mixture of two independent exponential random variables (hyper-exponential): with probability $1/3$ a mean 2 exponential and with probability $2/3$ a mean $1/2$ exponential. This distribution has mean 1 and squared coefficient of variation 2 agreeing with α_1 and v_1^2 as specified in (38).

The arrival process, $E_2(\cdot)$, is less variable. It is taken to be a renewal process with inter-arrival times that are Erlang 2 distributed this time with mean $1/4$. This is in agreement with α_2 and v_2^2 as specified in (38).

We consider two settings:

Single-class: Each queue has a dedicated (separate) server. This is a generalized Jackson network.

Multi-class: Queues 1 and 2 are served by the same server under a non-preemptive priority policy giving priority to queue 1. All other queues have their own server. Note that in this case the load on the server of queues 1 and 2 is $\lambda_1/\mu_1 + \lambda_2/\mu_2 \approx 0.97 < 1$. That is, it is quite heavily loaded but is still stable. Note in general having a load of less than unity does not immediately imply that the system is stable yet for this simple case it can be shown that stability holds under such a priority policy (cf. [4]).

Besides exemplifying the correctness of our theoretical results, the goal in this simulation setup is to illustrate that while the asymptotic variability parameters do not

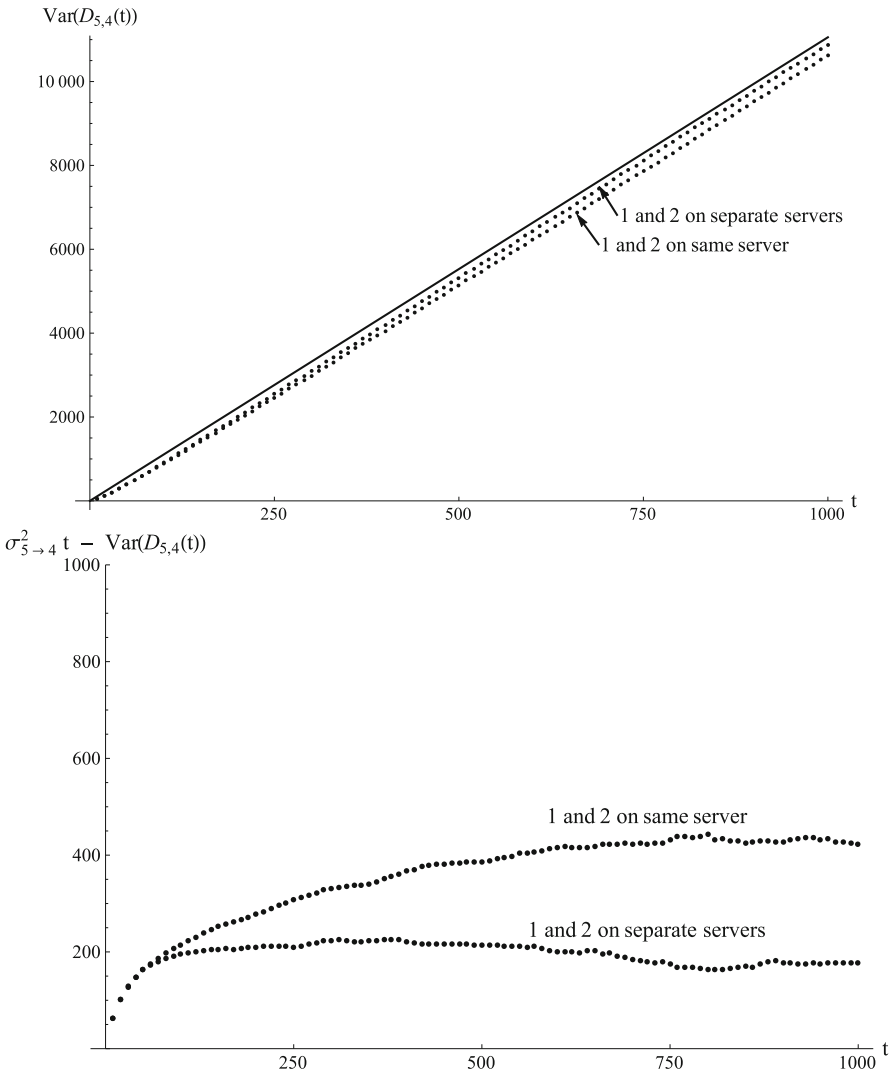

Fig. 2 Simulation estimates of $\mathrm{Var}\big(D_{5\to4}(t)\big)$ for two cases: single class (1 and 2 on separate servers) and multi-class (1 and 2 on same server with a priority policy). The top graph illustrates the variance curve estimates (dotted) vs. the solid line $\sigma^2_{5\to4}t$. The bottom graph shows the bias: $\sigma^2_{5\to4}t - \mathrm{Var}\big(D_{5\to4}(t)\big)$. As is illustrated, both systems have the same asymptotic variance for $D_{5\to4}(t)$, yet their variance curves differ for finite t

depend on service times and scheduling policies, the shape of the variance curve is in general influenced by such factors.

We ran 2×10^5 simulation runs of each case (single-class and multi-class) each for 1, 000 time units, starting at time $t = 0$ with the system empty[1]. We then estimated

[1] The simulation was carried out using a simulation package written in C++: PRONETSIM. See [32], Appendix A, for details about this software.

$\mathrm{Var}\big(D_{5\to 4}(t)\big)$ for each run over a grid of time points $t = 20, 40, 60, \ldots, 1000$, by taking the sample variance at each time point over 2×10^5 observations. Note that we purposely observe the flow $5 \to 4$ which is not directly adjacent to the multi-class server serving 1 and 2.

Our main theorem applied to this example implies that in both the single-class and multi-class case, for non-small t,

$$\mathrm{Var}\big(D_{5\to 4}(t)\big) \approx \sigma^2_{5\to 4} t = \frac{199}{18} t = 11.05\overline{5}\, t.$$

This is illustrated in Figure 2 (top) where we plot the variance curves versus the approximation $\sigma^2_{5\to 4} t$. To take a closer look at the effect of single-class vs. multi-class, we then plot the bias, $\sigma^2_{5\to 4} t - \mathrm{Var}\big(D_{5\to 4}(t)\big)$ in Figure 2 (bottom). It is indeed evident that different system characteristics yield different variance curves.

It is somewhat expected that the multi-class case will have a higher bias, since in this case the server of 1 and 2 is under a heavier load (0.97). Further, in that case one can expect more "bursts" on the flow $2 \to 4$ since queue 2 is served with low-priority. These bursts perhaps "propagate" to flow $4 \to 5$ and ultimately to the flow which we measure: $5 \to 4$. Nevertheless, such phenomena are not captured by the asymptotic quantities found in the current paper. It should be noted that in [22] second-order properties of this sort are explored for elementary queueing systems such as the stable M/G/1 queue. It is not clear how to extend such an investigation to networks.

7 Conclusion

Prior to this work, a rigorous analysis dealing with exact expressions for the asymptotic variability of flows was lacking in the literature. In this paper, we put forward easy computable expressions together with a simple diffusion limit theorem for the flows.

The queueing networks we considered in this paper are assumed to be open and stable. This stands in contrast with the more general case handled in [9] (where nodes are allowed to be either under-loaded, over-loaded or critical). It should be mentioned that our results easily carry over to the case where some nodes are over-loaded. In this case, the service times of over-loaded nodes contribute to the exogenous arrivals in a straightforward manner (see for example [18] for an early treatment of this idea). On the contrary, the case in which some nodes are critical is more challenging. In that case, the single-server queue was only recently handled with some difficulty in [1]. There the authors observed a BRAVO effect (Balancing Reduces Asymptotic Variance of Outputs). We do not handle this in the network context. Thus, the challenge of finding the asymptotic variability of flows in critical queueing networks remains.

Acknowledgements YN is supported by Australian Research Council (ARC) grant DP180101602. Part of the work was carried out, while WS was supported by an Ethel Raybould Visiting Fellowship to the University of Queensland. SBM is supported by the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) under grant number CE140100049. We thank an anonymous referee of this paper for their valuable comments and suggestions, particularly for highlighting the need for Assumption (A7). We also thank an anonymous referee of an earlier version of the paper for drawing our attention to the usefulness of the innovations method calculations.

References

1. Al Hanbali, A., Mandjes, M., Nazarathy, Y., Whitt, W.: The asymptotic variance of departures in critically loaded queues. Adv. Appl. Probab. **43**, 243–263 (2011)
2. Baccelli, F., Foss, S.: Ergodicity of Jackson-type queueing networks. Queueing Syst. **17**(1–2), 5–72 (1994)
3. Bean, N., Green, D., Taylor, P.G.: The output process of an MMPP/M/1 queue. J. Appl. Probab. **35**(4), 998–1002 (1998)
4. Bramson, M.: Stability of Queueing Networks. Springer, Berlin (2008)
5. Bramson, M., Dai, J.G.: Heavy traffic limits for some queueing networks. Ann. Appl. Probab. **11**(1), 49–90 (2001)
6. Bramson, M., D'Auria, B., Walton, N.: Proportional switching in first-in, first-out networks. Oper. Res. **65**(2), 496–513 (2016)
7. Burke, P.J.: The output of a queuing system. Oper. Res. **4**(6), 699–704 (1956)
8. Chen, H.: Fluid approximations and stability of multiclass queueing networks I: Work-conserving disciplines. Ann. Appl. Probab. **5**(3), 637–665 (1995)
9. Chen, H., Mandelbaum, A.: Stochastic discrete flow networks: Diffusion approximations and bottlenecks. Ann. Probab. **19**(4), 1463–1519 (1991)
10. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization. Springer, Berlin (2001)
11. Dai, J.G.: On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. Ann. Appl. Probab. **5**(1), 49–77 (1995)
12. Dai, J.G., Meyn, S.P.: Stability and convergence of moments for multiclass queueing networks via fluid limit models. IEEE Trans. Autom. Control **40**(11), 1889–1904 (1995)
13. Dai, J.G., Harrison, M.J.: Steady-state analysis of RBM in a rectangle: Numerical methods and a queueing application. Ann. Appl. Probab. **1**(1), 16–35 (1991)
14. Daley, D.J.: Queueing output processes. Adv. Appl. Probab. **8**, 395–415 (1976)
15. Disney, R.L., Kiessler, P.C.: Traffic Processes in Queueing Networks-A Markov Renewal Approach. The Johns Hopkins University Press, Cambridge (1987)
16. Disney, R.L., Konig, D.: Queueing networks: A survey of their random processes. SIAM Rev. **27**(3), 335–403 (1985)
17. Glynn, P. W.: Diffusion approximations. In: D.P. Heyman, M.J. Sobel (eds.), Handbooks in Operations Research, Vol 2, North-Holland, Amsterdam, pp. 145–198 (1990)
18. Goodman, J.B., Massey, W.: Non-ergodic Jackson network. J. Appl. Probab. **21**(4), 860–869 (1984)
19. Gut, A.: Probability: A Graduate Course. Springer, Berlin (2005)
20. Gut, A.: Stopped Random Walks: Limit Theorems and Applications. Springer, Berlin (2009)
21. Harrison, M.: Brownian models of queueing networks with heterogeneous customer populations. In Stochastic differential systems, stochastic control theory and applications, pp. 147–186. Springer, (1988)
22. Hautphenne, S., Kerner, Y., Nazarathy, Y., Taylor, P.: The intercept term of the asymptotic variance curve for some queueing output processes. Eur. J. Oper. Res. **242**(2), 455–464 (2015)
23. Jackson, J.R.: Jobshop-like queueing systems. Manage. Sci. **10**(1), 131–142 (1963)
24. Jacod, J., Shiryaev, A. N.: Limit theorems for stochastic processes, volume 288 of Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, second edition (2003)
25. Karlin, S., Taylor, H.M.: A First Course in Stochastic Processes. Academic Press, Cambridge (1975)
26. Kelly, F.: Reversibility and Stochastic Networks. Wiley, New York (1979)
27. Kemeny, J.G., Snell, J.L.: Finite Markov Chains. Springer, Berlin (1960)
28. Kim, S.: Modeling cross correlation in three-moment four-parameter decomposition approximation of queueing networks. Oper. Res. **59**(2), 480–497 (2011)
29. Kim, S., Muralidharan, R., O'Cinneide, C.A.: Taking account of correlations between streams in queueing network approximations. Queueing Syst. **49**(3), 261–281 (2005)
30. Kuehn, P.: Approximate analysis of general queuing networks by decomposition. IEEE Trans. Commun. **27**(1), 113–126 (1979)
31. Meyn, S.P.: Control Techniques for Complex Networks. Cambridge University Press, Cambridge (2008)

32. Nazarathy, Y.: On control of queueing networks and the asymptotic variance rate of outputs. PhD thesis, University of Haifa, (2008)
33. Nazarathy, Y., Weiss, G.: Positive Harris recurrence and diffusion scale analysis of a push pull queueing network. Perform. Eval. **67**(4), 201–217 (2010)
34. Reiman, M.I.: Open queueing networks in heavy traffic. Math. Oper. Res. **9**(3), 441–458 (1984)
35. Weiss, G.: Scheduling and Control of Queueing Networks. Institute of Mathematical Statistics Textbooks, Cambridge University Press (2021)
36. Whitt, W.: Performance of the queueing network analyzer. Bell Syst. Tech. J. **62**(9), 2817–2843 (1983)
37. Whitt, W.: Stochastic Process Limits. Springer, New York (2002)
38. Whitt, W., You, W.: Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function. Stochast. Syst. **8**(2), 143–165 (2018)
39. Williams, R.J.: Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. Queueing Syst. **30**, 27–88 (1998)
40. Zhang, F. (ed.): The Schur complement and its applications. Numerical Methods and Algorithms, vol. 4. Springer, New York (2005)